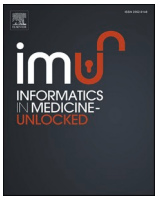
Informatics in Medicine Unlocked 46 (2024) 101454

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Application of glass box AI to large numbers of medical records for rapid response to future respiratory virus pandemics. Examples considering potential future high-fatality COVID strains and a potential avian influenza pandemic in humans

B. Robson [a,b,*], O.K. Baek [c]

[a] Ingine Inc., Cleveland, OH, USA
[b] The Dirac Foundation, Oxfordshire, UK
[c] Electronics and Telecommunications Research Institute, South Korea

ARTICLE INFO

Keywords:
Pandemics
Epidemics
Patient medicial records
Longtitudinal health records
Early response
Epidemiology

ABSTRACT

It is crucial to consider the consequences that new strains of respiratory viruses such as COVID-19 and avian influenza could have on humans. Possible future human-to-human transmission of avian influenza is of particular concern. As discussed, not all countries took a worst-case approach to COVID-19 at the outset, with regrettable outcomes. To better prepare, it is important to have access to as much information as possible, including digital patient records, and to use that information in a timely fashion so that appropriate actions can be taken early. A glass-box AI approach, complementary to current mainly black-box AI, can effectively manage uncertainty, missing data, and feature interactions in a probabilistic fashion. This approach can obtain standard epidemiological measures, discover unexpected demographic and clinical interactions in past data, and then apply them to small amounts of future data. As this concerns future response, this is primarily a review and position paper. It is emphasized that our results at both the quantitative and qualitative levels are based on models for future pandemics of unknown nature and possibly great severity and are not intended to be realistic. We may sometimes overemphasize severity, but that is a worst-case strategy. We do not consider all epidemiological modeling methods. Rather, this paper concerns how some simple, less variant measures from the first COVID-19 wave and more general qualitative information might be used in combination with analysis of rapidly updated patient records in the first few days of the first wave of a future pandemic.

1. Introduction

1.1. Background

After the influenza pandemic of the early 1900s and the success of vaccines against many infectious diseases, there was a period of complacency until the rise of AIDS fired interest in methods for rapid response to emerging epidemics [1]. We use the words epidemic and pandemic carefully but often interchangeably here. The word "epidemic" is traditionally applied to an outbreak in one country, while a pandemic is global. As modern people have learned of their cost, any serious communicable disease in one nation quickly becomes a more global concern. COVID-19 is obviously the most serious subsequent contender for a pandemic. According to the WHO and other sources, the four most dangerous transmissible diseases in humans around 2022 were COVID-19 and tuberculosis (probably promoted by AIDS) at 1.24 million and 1.13 million deaths respectively each year, and AIDS and mosquito-borne malaria at 0.63 million and 0.62 million, respectively each year. Influenza is estimated to cause some 0.4 million respiratory deaths each year on average across the world due to pneumonia and other respiratory symptoms. Somewhat like the case of HIV causative of AIDS, the COVID-19 outbreak in early 2020, for which the total number of deaths worldwide stands at about 7 million at the end of 2023, was not only an unexpected new strain of virus (in the case of COVID, of SARS - Severe Acute Respiratory Syndrome), but coronaviruses were a relatively new family as far as human infection was concerned [2–4]. Below, we tend to use the term "epidemic" rather than "pandemic" because the issue is no less serious when confined to a nation and, in the

* Corresponding author. Ingine Inc., Cleveland, OH, USA.
E-mail address: barryrobson@ingine.com (B. Robson).

https://doi.org/10.1016/j.imu.2024.101454
Received 14 December 2023; Received in revised form 27 January 2024; Accepted 27 January 2024
Available online 15 February 2024

critical early stages of a pandemic, almost inevitably is, though unfortunately typically only for a short period.

The importance of an even faster response to unfamiliar strains than that applied in the case of COVID-19 has been noted (e.g., Refs. [5,6]). In 2023, the World Health Organization identified over 50 COVID-19 variants, emphasizing the need to consider the possibility of different presentations of symptoms when diagnosing the causative agent. Author BR's interest in fast scientific response methods comes from the ability to respond with computational studies within a week [7] of the final corrected version of what was then called the Wuhan Seafood Market Coronavirus genome in January 2020 [8,9]. A peer-reviewed report of an extended study in February 2020 [10] soon followed this. These early responses noted the relation to SARS and a bat-hosted form and prediction of relatively well-conserved epitopes from the spike protein sequence for diagnostic, vaccine, and peptidomimetic drug leads. The 2002–2004 outbreak of SARS, caused by severe acute respiratory syndrome coronavirus, infected over 8000 people from 30 countries and territories. It caused at least 774 deaths worldwide. The information that was thereby unleashed for application to COVID-19 in 2020 included the possible benefit of natural agents such as emodin and derivatives [10]. Though by no means alone, even in these early days (e.g., Ref. [11]), these and the following series of papers [12–15] were undoubtedly rapid responses by previous criteria, including several novel aspects such as prediction of host cell sialic acid binding sites despite the absence of influenza-like neuraminidase [13].

Despite useful experience in both academic and commercial sectors with early response to epidemics such as AIDS and Mad Cow Disease [16–19], a difference in 2020 for coauthor BR was the access to an integrated system of biomedical knowledge gathering and prediction tools, including patient records. Integration of medical records for clinical and research applications was seen as lacking by US Presidential advisors in 2010 [20]. While some observers might argue that subsequent interoperability between medical record systems with different ontologies has not hugely improved, there are at least large archives of medical records becoming available (see Section 1.5). For that reason, the above approach has been developed over several years [10,12–15,21–32] but particularly in the last five [33–39], in which the emphasis has been on Big Data comprising millions of high-dimensional structured medical and health insurance records and collections of many million probabilistic "knowledge elements" [39]. These currently are being converted to the SCALA language primarily run on Amazon Web Services in SPARK environments [39]. These are relevant to Deep Learning as follows.

1.2. Focus and outline of the present paper

The authors wished to draw the AI attention of the community to the potential for epidemics of unprecedented severity, as with the case with avian influenza as described below, and not appear to compete one method against another and make judgements about merit that may not be appropriate in unprecedented circumstances. This did not emerge as an easy task. Ingine is a commercial developer of some of these tools discussed in Refs. [19,21–39] but the purpose of this present paper is not to promote them. Rather, it is to use some of their basic and intuitive features as a starting point for describing different AI methods and proposing potential future methods. This is possible and helpful because the goal of these tools was to develop a "universal exchange language" called Q-UEL that links a variety of knowledge-gathering, semantic web, and prediction sources, including different perspectives on probability [21]. The mathematical aspects of these tools are generally hidden "under the hood," and further and more detailed discussion of the theory here is not appropriate (except briefly in Methods Section 3.1) because it is based on Dirac notation and algebra. This is a standard well-known in physics and theoretical chemistry since the 1930s but less familiar outside of those fields. Consequently, Theory Section 2 is largely concerned with using some of the more obvious of these features to give an account of traditional medical probabilistic measures that can form the basis of glass box approaches. There, an apparent exception is a recommended approach using the partially summated zeta function to manage very sparse as well as plentiful data, which is compatible with the Dirac-based approach and appears as part of many of the above Dirac-based applications. However, this is arguably justified because it has an older history going back to use in bioinformatics applications, and versions are still widely used [38]. Moreover, a main theme of the present paper is to promote that or some similar method, because in the earliest and critical days of an epidemic, data related to the pathogen is likely to be sparse.

The present paper is also a position paper advocating for the inclusion of "glass box" approaches in AI as a means of responding early to emergent epidemics. These approaches stand in contrast to traditional logistic regression, Machine Learning, and Deep Learning methods, which are commonly referred to as "black boxes" due to the opaque nature of the patterns of weights that are learned. While some of these methods do attempt to add explanatory activity, such efforts are mostly retrospective in nature. By contrast, glass box approaches are explicit representations of knowledge, including probabilities where appropriate, from the outset. Moreover, these approaches are derived from real experimental probabilities. They can thus be used in conjunction with black box methods as complementary approaches. It is worth noting that the methods described in Section 1.1 are fundamentally glass box and are excellent examples of fast response methods. This review does emphasize the knowledge-gathering and prediction methods used. Still, their further purpose is to facilitate the use of probabilistic bidirectional general graphs for inference. The Dirac notation and algebra allow the use of relations that have verbal, rather than just conditional, force. This notation provides a convenient way to express and package knowledge and probabilities as the basis of automated reasoning and predictions. Additionally, while some of the results presented in Results Section 4 are based on Dirac-based methods, it is reasonable to expect that many of the findings will be applicable to glass box approaches of a similar but not identical nature.

A great deal of the information reported here, and several of the approaches described, may appear to be like the use of familiar traditional statistical tools. To any extent that this might be argued to be so, it would still be valid because a major purpose of the present paper is to alert the AI community particularly to the danger of avian influenza. It is also to highlight the features in rapidly updated medical records on which both black box and glass box machine learning methods may focus in any such emerging epidemic. This is to reduce dimensionality and increase response time, and it may be identified with *feature selection* or *feature extraction*, i.e., a phase in many AI methods to determine what attributes are relevant and non-redundant, with the aim of reducing dimensionality prior to any study. In addition, there is of course little merit in producing intended useful information in a form that appears alien.

However, there are significant differences in the tools used in the present paper. First, the above-mentioned universal exchange language Q-UEL [21] was used to perform automated surfing and unstructured data mining of natural language text, and of tabular information, on the Internet [19,30,31]. This is important because it was invaluable in a rapid response to COVID-19 [10–15,19], in which papers many examples and detailed descriptions may be found; it may reasonably be assumed that some such kind of approach will be important in the early stage of future epidemics. The information automatically collected here (some already the result of traditional statistical analyses) was where necessary collated, tidied, and statistically analyzed or further analyzed by Q-UEL applications. In several respects this gives the present paper some of the character of meta-analysis and systematic review. It yielded Q-UEL's XML-like knowledge representation tags like the examples in Methods Section 3 below and presented extensively in Refs. [10–15,19]. These represent a standard canonical form that may be retained for future use in a Knowledge Representation Store. The contents of these

tags can be displayed directly, or used in automated reasoning and in the preparation of reports and research papers. However, the medical features that such knowledge-gathering methods predict in the present paper as likely to be relevant to a new respiratory virus epidemic, so this information might also themselves save some time. Second, "knowledge gathering methods" also include strategies that appear to involve, and to a considerable extent do resemble, traditional data mining of structured data, notably medical records in this case. Some kind of data mining is a common first step in a glass box approach, analogous to the training step in neural net and logistic regression methods. However, this approach when used here differs from traditional data mining by specializing in high dimensional data and in the management of sparse data that is an inevitable consequence of it [28]. The major difference, however, is in the second step. Any good data mining approach is already useful in discovering unknown positive and negative associations, and in quantifying those already known. The important further step is in using the quantitative information found to make predictions and help make decisions. The method for that used here [33,39] employs less familiar algorithms related to quantum mechanical basis discussed above, but still makes visible familiar probabilities and odds etc. throughout. Like the above unsupervised data mining, it also performs its own data mining in order to get the additional information needed for its much less usual "quantum" approach (so in many cases the first step is not required, except for providing guidance as to a supervising query). For the analysis of medical records, both of the present authors use data mining as well as black box AI. However, author BR favors the above glass box prediction approaches for maintaining comparable predictive performance while using and generating probabilities and probabilistic measures of interest to epidemiology and Evidence Based Medicine as discussed later below.

It remains that Deep Learning and new related methods are, of course, powerful tools and of great interest to both authors and the AI Institute of ETRI. The healthcare concerns of ETRI in the present government research grant include developing efficient and scalable methods for the analysis of longitudinal health records, including for consideration of the reinfections by strains of COVID-19 virus [40,41] as well as entirely new pandemics. Outside our present collaboration, computational methods of many kinds have been applied to predict new variants by many researchers (e.g., Refs. [42,43]). Deep Learning has already played a prominent role in COVID-19, as has been reviewed elsewhere [44]. The reason why the glass box strategy is of interest to both authors' approaches is not simply that glass box approaches are complementary by adding explicability to black box methods. It also has many of the characteristics of feature extraction discussed above. These capabilities are also discussed below. The main concerns and targets remain the epidemics themselves, as follows.

1.3. COVID-19 and avian influenza

This paper focuses on possible future pandemics from (i) an aggressive new strain of COVID and (ii) a possible widespread transmission of avian influenza between humans in part because of their plausibility and the seriousness of the impact on social structure, as discussed below. Diseases like influenza and COVID-19 are well known to be *zoonotic diseases* that come from animals within our current lifetimes [1]. Virus epidemics start with an animal-to-human jump, then require a human-to-human jump that is typically much less easy: it is the last barrier in the case of avian influenza that awaits the appropriate mutations or simply the reassortment of genomes discussed shortly below. Both flu and COVID are due to respiratory viruses with similar dependencies on conditional factors such as comorbidities and the relevance of various clinical laboratory tests. Both belong to the realm of Riboviria (RNA viruses), and both are single-stranded "chromosomes" but differ in being positive and negative sense. Both have virions (virus particles) roughly 80–120 nm in diameter. However, influenza A has a smaller genome, 13.5 kb, compared with about 30 kb for COVID-19. Despite that, the flu genome is much more obviously segmented into modular sections, representing a primitive form of "sexual reproduction" and fast evolution in which different strains, including porcine, human, and bird host strains, can practice crossing-over of genomic material called *reassortment*. The less well-developed capability in coronaviruses does not seem to hinder their remarkable ability to jump host species. Both have lipid envelopes with spike-like proteins, and both can bind host cell sialic acids. The COVID-19 virus differs from influenza (and many coronaviruses) by lacking neuraminidase to reverse that process. Perhaps contrary to popular belief, many researchers have considered that both flu and COVID-19 are probably primarily upper respiratory tract infections (RTI). Most influenza infections affect the upper tract but can be an upper or lower RTI, commonly with inflammation of the upper respiratory tree and trachea. Lower RTIs tend to last longer, can be more serious, and have been faily commonly interpreted as an extension of infection from the upper airways. In COVID-19, upper respiratory symptoms present even more frequently than in influenza virus infections, but lower tract infection is indicated in 30% of COVID cases. Respiratory symptoms such as cough, sputum production, and breathing difficulties are largely due to lower tract infection. Often, serious respiratory symptoms in flu and COVID can indicate secondary infections such as pneumonia caused by bacteria and, on occasion, perhaps fungal sources.

At the time of first writing, the imminent potential pandemic of concern to epidemiologists was the case involving a new strain of porcine influenza that, on early evidence, could represent a human-to-human rather than simply a pig-to-human jump [45]. However, compared with a fatality rate of 0.1%–10% for the first 100 million cases of COVID-19 varying with country, and even 10–20% of those who died infected by Spanish flu, avian influenzas A(H5N1), A(H5N6), and A (H7N9) have had fatality rates of up to 50%–60% or more [19]. Amongst many bird species, H5N1 is highly contagious, and the fatality rate, specifically meaning the fraction of infected individuals that die, the rate has been about 100% [46]. The R_0 factor (spreading power) could be huge in humans. The average incubation period of bird flu H5N1 is two to five days, though it can last up to 17 days. For H7N9, the average incubation period is five days and can last up to 10 days. Both viruses have a longer incubation time than seasonal influenza [47]. "As of February 23, 2023, a total of 240 cases of human infection with avian influenza A(H5N1) virus have been reported from four countries within the Western Pacific Region since January 2003 …. Of these cases, 135 were fatal, resulting in a case fatality rate (CFR) of 56%" according to the WHO February 2023 [48]. Koopmans (who investigated the origins of the COVID-19 pandemic for the WHO) – warned: "We are playing with fire." [49]. By mid-2020, there were probably 10^{27} SARS-COV-2 RNA molecules in the world, a massive reservoir for accepted mutations [19]. Still, while Earth's human population is about 8 billion, there are between 50 and 430 billion flying birds. Researchers have already created hybrid viruses by mixing genes from H5N1 and the H1N1 strain behind the 2009 swine flu pandemic and showed that some of the hybrids can spread through the air between guinea pigs [50].

1.4. Consequences of a serious pandemic

A major motivation for the present paper is the serious consequences of human-to-human transmission of avian influenza. A healthcare system is likely to be the first organization overwhelmed in a nation in the event of a serious epidemic, and this is the primary reason for a lockdown, as in COVID-19. The capacity of a nation's healthcare system is conveniently, if not ideally, indicated by the number of hospital beds and the percentage normally unoccupied to manage an epidemic wave. Since 2010, average NHS UK bed occupancy has consistently surpassed 85%. In the US, around 2019, the occupancy rate of hospitals stood at 64.4 percent. Korea, Japan, Russia, and Germany. Japan has around 1.26 hospital beds per 100 population. The United States reported just 0.28 hospital beds per 100 population, and the UK about 0.24. While a

low number of hospital beds in a nation can reflect a high standard of preventative, basic, and ambulatory patient care, this does not help in the case of a novel epidemic with a high incidence and prevalence of severe forms of the disease. The Omicron BA.1 COVID-19 variant peaked in London at 8.79% of the population infected and in the UK as a whole at 2.69% on January 7th, 2022.

In global estimates, 20% of infected people with COVID suffered a severe or fatal form of the disease, giving about 1.76% of people with serious COVID-19 in London and 0.54% in the UK overall. The popular UK press considered that the NHS was strained even before the arrival of COVID-19. At the end of 2022, it was still teetering on the very edge of collapse. In the 1918–1919 influenza pandemic, an estimated 33% of the world's population was infected. Dengue, Ebola, and the bubonic plague of 1347–1351 remind us that high fatality rates from emergent epidemics are plausible.

A statement that is usually attributed to the US Centers for Disease Control around 2000 is that between 15% and 35% of the U.S. population could be affected by an influenza pandemic. The economic impact could range between $71.3 and $166.5 billion. That means something like $300 billion or more in 2023, but the important question is, why choose "35% of the U.S. population?" Many have in the past considered that much more than 35% infection would mean "game over." The truth, however, is that no publicized scientific study has looked at whether a pandemic with high mortality could cause social collapse [51]. It seems plausible that as happened in 1918, as infections and mortalities increased, cities around the world could collapse [52]. David Brin's 1985 post-apocalyptic fiction book "The Postman" and the later movie [53] describe the future US as an essentially medieval system, but without a national central authority, 13 years after unspecified apocalyptic events followed by plagues, and take the US Postal Service as the symbol of social cohesion, and communication loss as a key feature of social breakdown.

1.5. Previous work on medical records

Review will be given in the Sections dedicated to Theory, Methods, and Results as appropriate. In this Section, the focus is confined to the intended starting point, i.e., medical records. They are the key sources for providing detailed information about epidemics and their effects, and they are not without historical precedent. Examples are found in Ancient Egyptian papyri, the writings of Hippocrates, and letters to him asking for help regarding the help with the great plague in Athens. That plague had symptoms well described by historian Thucydides (*circa* 460 - 404 BCE). Such comprehensive descriptions from ancient history could give us some indication of a possible fatality rate if appearing in an epidemic today. The modern analysis of the above plague is so far less clear, but most often in favor of smallpox or typhus, and more recently, Ebola. The Regenstrief Institute in Indianapolis is usually credited with creating the first digital patient records in 1972. However, in 1962, IBM installed an electronic medical record system at Akron Children's Hospital in Ohio. The UK's National Health Service did not adopt EHRs until 2005, despite being an integrated system from a management perspective. Their goal was to have a widely used centralized system by 2010.

The Longitudinal Health Record (LHR) is a complete patient record that combines data from various sources throughout the healthcare system. A cradle-to-grave version of the LHR documents the patient's entire medical history. A large number of LHRs that are updated in real-time and accessible in de-identified and well-structured forms by medical researchers would be even more valuable for learning from and responding to epidemics. Early observations that the COVID-19 outbreak was indeed a form of SARS [7,10] were important because they allowed a significant amount of information to be applied to COVID-19 in 2020. Fatality rates can, in less industrialized nations, be a matter of pencil-and-paper survey and census. In others, computer-based surveillance systems and/or smartphone apps were important, but access to a large number of digital patient records has the potential to provide more than just basic epidemiological measures. Clinical data can be combined with molecular studies of the virus, as described in Section 1.2. This combination can provide a better understanding of the interaction between the pathogen and the host on a practical, statistical, and nationwide scale. Truveta has access to 100 million digital patient records in the US [54]. The NHS and the South Korean Electronics and Telecommunications Research Institute also have large sets of data available. With access to all of these data sets, Ingine's approach to understanding the interaction between the pathogen and the host can be a powerful tool at a critical time [12–15]. Truveta cites several papers on their website about ways in which many patient records can shed light on the clinical aspects of COVID-19 [54], e.g., Ref. [55]. These kinds of efforts should perhaps be distinguished from recent efforts in DL for the identification and prediction of new strains of COVID-19, which put much less emphasis on patient records. However, they can still play a role in terms of determining or predicting fatality and morbidity rates as influenced by other demographic and clinical factors [56]. For the rapid application of AI to medical records generally and longitudinal, cradle-to-grave medical records particularly, the present authors, along with Ingine, have developed a structured data ontology [56].

2. Theory

2.1. Ontology

There are some core concepts to consider not only because they can be technical but also because those that are familiar are not always as trivial as they first appear. The notion of ontology is important to the organization of medical records discussed above in Section 1.5, originally so that they are comprehensible by humans and later also to facilitate analysis. Rigor in notation is especially required if the semantic structures implied are to be efficiently manipulated by automated reasoning beyond basic statistics, e.g., to extend the Worldwide and Semantic Web to a "Thinking Web" [25].

By ontology, here we mean the use of some attributes, states, events, factors, descriptive parameters, observations, or measurement A:=a, which is observable and countable, where A is the *attribute name,* and a is a specific value of it *distinguishable* from other possible values. ':=' may be considered as a *metadata operator* such that what lies to its left has some force as an explanation, and what lies to its right is a more specific case or example. Particularly for analysis of patient records as data, we mean such as 'congestive heart failure':=yes, age(years):=35, or 'systolic BP (mmHg)':='160–169' [21], although ontology can also mean a graph structure such as A:=(B:=b, C:=D:=d) in which, in effect, one or more attributes can become the values of another [25,56]. A simple example is 'blood pressure':=('systolic (mmHg)':=140, 'diastolic (mmHg)':=70. In structured data such as a comma-separated value (CSV) file, the example A:=(B:=b, C:=D:=d) could be a header. However, that would mean that there is a pair of coupled values, b, d, as a datum in each row below it, so if this explicit approach is taken, one would be more likely to see structures such as A:=B:=C:=c where there is just one value, c, in each row below it. In either kind of case, for a record overall one may still say there is a *local significance* (often assumed) such that on data analysis the joint attribute (A:=(B:=b, C:=D:=d)) has a logical true or false value (A:=(B:=b, C:=D = d)), or a probability P(A:=(B:=b, C:=:=d)) based on counting n(A:=(B:=b, C:=D:=d))/N observations where N is the total amount of data given by n(A:=(B, C:=D)), excluding where values are unknown [39]; that is even if a much larger graph can model the overall system of interest called an inference net, in which case A:=(B:=b, C:=D:=d) is an example of a subgraph.

When used in prediction mode, typically meaning in *inference nets* [26,34], methods based on the above can be considered as "glass box" approaches because of the local significance and notation that involves probability values (or other probabilistic measures) with explanatory

annotation. Inference nets, in general, are most important when they make estimates of complicated probabilities (with many attributes) because there is inadequate data to perform such counting directly. They do so by multiplication of simpler probabilities that can be determined by counting, which assumes independence between attributes in each probability, which is indicated by the overall graph structure. This assumption of independence happens whether one is discussing the use of an overall graph structure, which is a Directed Acyclic Graph, i.e., Joshua Pearl's Bayes' Net reviewed in Refs. [26,34], or the Bidirectional General Graph [34]. They are all estimates unless very attribute is assigned a relationship with every other attribute collectively, meaning that there is a joint probability, say P(A:=a, B:=b, C:=c, …) accessible to counting, which is, in that sense, exact and replaces the need to have any graphic representation, and when required is convertible to a conditional probability (often attributed to de Moivre 1667–1754) such as P (A:=a |B:=b, C:=c, …) = P(A:=a, B:=b, C:=c, …)/P(B:=b, C:=c, …).

At first glance, the essential features of most machine learning methods related to the neural net, such as Deep Learning, seem potentially accessible to the above considerations and even appear relatively simple. They take the input variables and the above linear combination equation of $Z = W_0 + W_1X_1 + W_2X_2 + \ldots + W_nX_n$ to compute the output or the predicted values for each node in the next layer. The weights associated with them usually have no individual intrinsic ontology or meaning. The weights are optimized to fit the data by calculating a loss or the error term, which is the deviation of the actual values from the predicted values, and which the algorithm minimizes. The above formula for Z implies familiar high-school linear regression, and it is evident that such methods are essentially an assembly of regression-based operations integrated into an overall graph structure, typically one with progressive layers of nodes. The output of each regression-like step is the input, usually processed by a transformation akin to normalization to another. Despite all that, the popular use of a sigmoidal function has similarities to the conversion of information to a probability, and the network suggests opportunities for an ontology.

Nonetheless, the approach is considered a "black box" because only the overall structure can be associated with meaning, and weights cannot typically be identified with probabilities in the statistical sense. This lack of statistical relevance is not least because the way in which the minimization process organizes the weights can have multiple solutions, many of which can frequently be effective, While a large system of the above kind can be considered as a *tabula rasa* that can learn anything given time and computational resource, speed is essential in applications of such methods to epidemics such as new strains of influenza or COVID-19 (e.g., Refs. [45,57]) where it is also evident that the first layer seen as an input layer benefits from a good choice of well-organized and appropriate technology when applied to epidemics. The structured data ontology proposed in Ref. [56], like many others, did not include fine details for the appearance of the effect of epidemics and the progress of an infected patient, but it was designed to be extensible.

2.2. Example ontology for analysis of future respiratory virus epidemics

Ref [19] proposed an example ontology that could be included in any effort of the above kind. This ontology comprised several patient states: E for Exposed, I for Infected and Incubating, P for Pivotal, Prodromal, and Propagative, S for Symptomatic, A for Amelioration, R for Recovery, N for Normalized, C for Complications, and D for Death. The P state was envisaged as leading to a logical exclusive OR branchpoint, with four outcomes, as follows.

(i) Normalization alone is involved in asymptomatic outcomes
(ii) Symptomatic state and Normalization in recovery
(iii) Symptomatic state and Complications and Normalization in hospitalization
(iv) Symptomatic state and Complications resulting in Death

The primary purpose of the above definitions of patient state as E, I, P, S, A, R, N, C, D is to assign probability weights to patient states as indicated above, and hence, overall path based on patient histories. The present author, BR, using Ingine technology, and colleagues used this type of approach for COVID-19. However, it was noted that recalibration was required for each strain [19]. That requirement is almost the definition of a "variant of concern" [19].

2.3. Critical E-I-P-S phases for respiratory virus epidemics

Not least from the individual patient's point of view, phases A, R, N, C, and D are obviously of concern once infection is established. Still, for fast response to emerging epidemics, the E-I-P-S phases are crucial in the interest of populations as a whole and in preparing healthcare systems. Rapid updating of medical records combined with rapid application of analytic and predictive methods remains important whenever possible. Still, the E-I-P-S phases of a future epidemic could be very short and almost asymptomatic. This timing is reflected by the importance of "lockdowns" and later exposure "apps" for the E phase, widely available diagnostic tests for the I phase, and reporting and digital surveillance for population studies for the P and S phases.

Unfortunately, appropriate use of all the above requires some prior knowledge, and for a new class of causative agent the propagative (infectious) P phase might exist much for of the E-I-P-S period. For COVID-19, the symptomatic S phase could appear 2–14 days after exposure. Still, children and adults with common influenza can infect others entering the P phase beginning about one day before the S phase and up to seven days after the symptoms resolve as the S phase. Patients with weakened immune systems could remain contagious for COVID and influenza for several weeks. Fortunately, when there have been related epidemics, much information from them is enormously useful. It can provide elements of knowledge that can be used in predictions later [11–15]. For example, once COVID-19 was accepted to be a form of SARS, it suggested several possible therapeutic agents [12,13]. On average, the incubation phase I for common influenza lasts 1–4 days, and on average, people start to develop phase S flu symptoms about two days after exposure. If illness is not apparent too soon, and certainly if serious morbidity or death does not come too soon, this facilitates the spread of the pathogen, which is the essential character of its evolutionary survival. A major epidemiological problem for the AIDS epidemic was that the first S phase signs can sometimes take months to years for any symptoms to appear. Although it is not guaranteed, the more general trend for pathogens, including influenza and COVID-19, is that S-phase symptoms in successive outbreaks have become milder and less obvious for longer. This delay aids any subsequent action by "buying time." Still, it complicates detection and speaks for the importance of widely available diagnostic tests, highlighting the fact that after a positive COVID-19 test result, it could, in many cases, take up to 10 days for the infection to develop.

So far, information regarding avian influenza must come from bird-to-bird and bird-to-human transmission since. Fortunately, reliable evidence of transmission between humans has been, at best, extremely rare (at the time of writing). For chicken-to-chicken transmission, the strains called highly pathogenic avian influenza (HPAI) viruses cause severe disease with high fatality. HPAI A(H5) or A(H7) virus infections affect multiple internal organs with fatality rates of up to 90%–100% within 48 h. However, ducks can be infected without any symptoms. The less pathogenic avian influenza strains (LPAI) are often close to asymptomatic in chickens but seem capable of mutating to HPAI. On average, human symptoms of avian influenza acquired from birds have appeared 3–5 days after infection. The average incubation period of bird flu H5N1 is 2–5 days, though it can last up to 17 days, and for H7N9, the average incubation period is 5–10 days [47]. This time frame is still more than ample for a seriously high R_0 (spreading factor). The main symptoms for humans infected by birds include a high temperature, aching muscles, headache, cough, and shortness of breath. Other symptoms may include

diarrhea, sickness, stomach pain, chest pain, bleeding from the nose and gums, and conjunctivitis. One is tempted to feel that an analogous influenza cannot be excluded from being a plausible diagnosis for the Plague of Athens described in Introduction Section 1.5.

2.4. Theoretical aspects of "black box" approaches in trial studies

An important part of a response to a new epidemic is the need for clinical trials [58]. New diagnostics, vaccines, and therapeutic drugs will almost always be needed. Even before that, as with COVID-19, existing therapies will be investigated for repurposing [10], followed by human trials. For a fast response, a more efficient trial design is needed. It has not gone unnoticed that clinical trials [58] and the future of clinical development generally are "on the verge of a major transformation due to convergence of large new digital data sources [and] computing power to identify clinically meaningful patterns in the data using efficient artificial intelligence" [59]. The review by Shah et al. [59] refers to publicly available biomedical and clinical trial data sets. Still, the focus is largely (but not solely) on "black box" approaches of Machine Learning ML including Deep Learning DL, and the architectures needed for them in the clinical context. Previously, it has often been said that the drawback of trials that utilize historical controls, which is part of the implicit role of retrospective cohort studies from many medical records, is that they are non-randomized studies, a non-randomized comparison thus subject to considerable bias, requiring additional assumptions when making group comparison. This presumably was, at least in part, because the historical data frequently related to another, outdated, and perhaps specific group. Still, it is also possible that this is simply because the analysis will show that very little in medical data is truly random and that many strong associations can be unexpected. When historical data are very reliable and well-documented, and other disease and treatment conditions have not changed since the historical data was obtained, then they can be considered. However, Shah et al. [59] also note that despite many propositions for the use of ML to accelerate medical research, very few successful use cases have emerged. They conclude that limited successes have been attributed to insufficient time elapsed since the introduction of relevant technologies and deficiency of current computer science DL and related ML models to generalize more complex and realistic medical data sets and tasks, nascent regulations, and ethical and legal concerns about data sharing. These are primarily practical matters, but they also blame the paucity of large numbers of high-quality labeled data, which we argued above to be essentially a theoretical matter and one that is particularly important in the epidemic context because speed is required for application of such methods to epidemics such as influenza or COVID-19, (e.g., Refs. [45,57]).

It is here where additional theoretical issues arise. In "computational Analysis" in 1, "Use cases of artificial intelligence, computer vision, and machine learning in clinical development," Shah et al. [59] give prominence to "feature extraction" even prior to neural networks, analytics, and visualization. This approach is partly a choice of suitable ontology that can be seen as a preparative step for black and glass box techniques. Still, importantly, it reduces the time-consuming high dimensionality of medical data by (i) identifying and removing irrelevant attributes and (ii) reducing the number of attributes that are essentially the same thing, i.e., all attributes except one are redundant information, and one will suffice. A challenge is that these two issues reflect very different *association constants*. The *atomic association constant* K(A:=a; B:=b; C:=c; …) = P(A:=a; B:=b; C:=c; …)/P(A:=a) P(B:=b) P(C:=c), for example, is close to 1 for irrelevant attributes that come together at random, but much greater than 1 for those that associate so closely that essentially mutually redundant. While seen as preparation of input for a black box (or glass box) method, these are inherently visible probabilistic matters that form a natural part of a glass box approach [60,61]. They imply an important glass box part of a black box approach. A worked example for a prediction and estimation of the Likelihood Ratio LR and related measure called DOR* is given in 8.1 of Ref [61]. The automatic construction of large odds inference nets by the specification of interdependent and independent factors and using the above information-theoretic ideas "under the hood" is a function of the DiracSmash module of the Ingine platform [33,39].

2.5. Theoretical aspects of traditional "glass-box" approaches in trial studies

There is also a broad range of relevant and more traditional statistical approaches to cohort studies and the design of clinical trials [58, 60–63] that are glass boxes in the sense that they make visible probabilities, odds, and other related probabilistic measures and statistics that have a significant history [64,65]. Currently, much focus is on the consequences of the new epidemic for patients with comorbidities such as diabetes [66], for estimating hospital requirements and for clinical trial design. Since it is recognized that correct judgment in the first few days of COVID-19 was critical [67,68], much can be learned from approaches from the methods of data collection and analysis used there [69–73]. The power of data collection is likely to be greatly extended by the emerging field of the Internet of Things (IoT), which is constantly increasing, providing a plethora of potential integration across many sectors, including healthcare [74]. Analysis and learning from sensors were also reviewed in Ref. [59].

Many of the issues discussed here traditionally lie in the domain of evidence-based medicine (EBM), which is primarily intended for use by physicians [75]. Still, the increasingly available new option is to see the application of glass box and black box methods, which essentially represent a new class of *retrospective cohort studies* made possible by the increasing accessibility of large numbers of digital patient records. One important feature of the traditional methods, the 2×2 structure, has been inherited. Like traditional cohort studies, whether retrospective or prospective, a major type of clinical trial called a *factorial trial* can be considered as having a basic building block of a 2×2 contingency structure [60]. Minimally, these might be patients with and without the disease of interest.

Each of these groups is divided into two halves, one receiving the new therapy and one a placebo or previous gold standard treatment. Trials in which a gold standard replaces the placebo are often referred to as *non-inferiority trials* [60]. In the common kind of trial called a *placebo-controlled trial*, only patients with the disease are enrolled and divided into a group receiving the therapy and the other the placebo [60]; evidently, this a special limiting case of the 2×2 trial with the risk for healthy patients in mind, but it is not a restriction for relevant retrospective cohort studies using medical records because any adverse events have already occurred. Obviously, if a healthy patient with known demographic and clinical characteristics had shown an adverse response to a gold standard drug, the same could apply to a patient with the disease. Another kind of experimental factorial trial with a 2×2 structure avoids risk to healthy patients and instead puts the emphasis on studying the effect of two or more interventions applied alone or in combination, say four intervention groups are defined based on whether they receive interventions A only, B only, both A and B, or neither A or B [60]. However, these same combinations can be usefully explored by purely computational analysis of many patient records. High-performance computational analysis of patient records has the obvious advantage that the basic 2×2 structure can be further split into

groups with different demographic and clinical factors. Trials can be considered for more than two interventions, resulting in a *multifactorial trial* 2 × 2 x 2 × … structure as the general case [62]. As an experimental cohort study or clinical trial, there would be the downside that candidate recruitment is more complex in the more complex factorial trials and can decrease accrual rates. Candidates must satisfy the criteria for treatment with each intervention with no contraindications to any treatment combinations and consent to all the interventions and combinations of them. Meanwhile, adherence to protocol is harder due to the multiple interventions and greater burden on all stakeholders. However, in such cases, the purely computational retrospective cohort studies thus become of even greater importance to justify trial structure and minimize risks.

2.6. The Diagnostic Odds Ratio DOR and its useful development as DOR*

Predictive Odds (POs), Likelihood Ratios (LRs), and Diagnostic Ratios (DORs) might be considered the simplest glass box measures that can also be used in a predictive way. The DOR* method described below as an extension of DOR already has eight probabilities in it, enough to see it as already taking on the character of a small glass box inference net. POs, LRs, and DORs are characteristic of EBM [75], which, as noted above, was primarily intended for use by physicians. Comparative Effectiveness Research (CER) [76] was originally perceived more as the counterpart of EBM in the USA, differentiated by matters such as emphasizing comparison of the merits of many interventions (therapies) rather than just two, but such criticisms may not have been intended too seriously because EBM's Problem-Intervention-Comparison-Outcome method (PICO) [75] is obviously extensible from two to many. It is fairer to state that CER is somewhat more focused on healthcare administration and strategy. Some authors consider the DOR to be the single most important indicator of test performance (e.g., Ref. [77]), though it does not usually capture in a single measure the relative merits of different interventions for good and bad outcomes. To the authors' knowledge, neither EBM nor CER makes mention of something akin to a DOR* that does that; it is a recently defined measure [62] explained below that appears novel or at least previously rarely used. We consider this in terms of multiple 2 × 2 x 2 data substructures where an appropriate measure of risk summarizes each one. The notion is simple: the well-known diagnostic odds ratio (DOR) [58,62] for a specified disease conditional on a favorable outcome is divided by the corresponding DOR conditional on an unfavorable outcome. At first glance, the obvious prediction to make is about whether the outcome is good rather than poor. A typical measure for such predictive purposes in clinical practice is the Likelihood Ratio LR [58,62].

$$\mathrm{LR}(\text{conditions}|\text{good outcome} : \text{poor outcome}) = \frac{\mathrm{P}(\text{conditions}|\text{good outcome})}{\mathrm{P}(\text{conditions }|\text{poor outcome})} \tag{1}$$

An outcome is said to be good if LR > 1 and poor otherwise. The larger the LR, the greater the odds that the outcome will be good.

Recall that a conditional probability (i.e., including the vertical bar '|' that may be interpreted as logical IF) is, for example, P(A:=a |B:=b, C:=c, …) = P(A:=a, B:=b, C:=c, …)/P(A:=a) P(B:=b) P(C:=c), …. That ':=' notation still applies, and it is what follows below, and 'conditions' may include 'bleeding peptic ulcer':=yes and a variety of demographic and clinical factors in similar format. Hence, note that the conditional bar is not confined to following the first attribute. We may have, for example, P(A:=a, B:=b, C:=c |D:=d, E:=e …) = P(A:=a, B:=b, C:=c, D:=d, E:=e …)/P(D:=d) P(E:=e) … Also, 'good outcome' may more specifically be, for example, 'first serum lactate dehydrogenase':=high and 'low serum lactate dehydrogenase':=low while 'bad outcome' might then be 'first serum lactate dehydrogenase':=high and 'low serum lactate dehydrogenase':=high [62]. The LR is independent of the prior odds PrO=P(good outcome)/P(poor outcome) in a population. Compare the Predictive Odds PO, normally favored as the appropriate measure for studying the health of populations.

$$\mathrm{PO}(\text{good outcome} : \text{poor outcome}|\text{conditions}) = \frac{\mathrm{P}(\text{good outcome}|\text{conditions})}{\mathrm{P}(\text{poor outcome}|\text{conditions})} \tag{2}$$

Recall that The LR is the PO for a patient compared with the PO that would be expected on a chance basis, which is reasonably taken as the prior odds PrO (i.e., LR = PO/PrO).

It is of interest in the present review that the conditions in the above equations will include a new COVID-19 strain or variant or an emergent human-to-human avian influenza. The best biomarker for success must be explored in the event of a new pandemic. Something of an immunological or genome sequencing nature would certainly be reasonable. Still, a common "best bet" generally suitable choice is the use of serum levels of aldehyde dehydrogenase ADH, which is well known to be generally raised in response to tissue damage and cancers and (roughly speaking) is a kind of "something is wrong" signal. If initially high and normal at the end of therapy, one can reasonably presume that not only has the treatment been effective but that there are no damaging side effects. In the case of other comorbidities being present, of course, the result of the last reading being high will be ambiguous, but that can be explored by analysis of many records as described below. To facilitate its use, low ADH is rare and confined to certain conditions. It is well known that, as might be expected, high ADH does occur in COVID-19 infection, and there are useful isoforms that can make finer distinctions; for example, there is an inverse relationship between the mitochondrial isoform ALDH2 *rs671* and antibody production.

The final measure of interest here is not the PO or LR, largely because it is not directly suited to a 2 × 2 x 2 contingency table and data structure. They are also somewhat inadequately robust for Real World Data, especially if the initial set of conditions specified is large. The DOR focuses on the disease, and this has advantages. The DOR is mainly used in an analytical mode in this study as follows. However, its predictive power can be expressed as a DOR expressed in terms of true and false positives and true and false negatives, TP, FP, TN, and FP, respectively.

$$\mathrm{DOR} = \frac{\mathrm{n}[\text{target, conditions}]\ \mathrm{n}[\text{not(target), not(conditions)}]}{\mathrm{n}[\text{target, not(conditions)}]\ \mathrm{n}[\text{not(target), conditions)}]} \tag{3}$$

This approach is classical, and n[ ] is the number of observations, i.e., a count. In this case, the "target" is, for example, congestive heart failure, kidney disease, or bleeding peptic ulcer. "Conditions" means the other attributes found, believed, or hypothesized to impact the target seen on a record, representing a joint event. We replace "target" with a new COVID-19 variant or bird influenza in an epidemic in humans. In Eqn. (3) "not( )" means the negation of the attribute or set of attributes stated, say, X, such that P(not(X)) = 1 – P(X). The DOR has the advantage that it is self-normalizing and would give the value 1 if the disease and conditions, and the absence of infection and absence of the conditions, were all independent (randomly associated).

The natural logarithm of the DOR, called LDOR, is then natural as a particular measure of information, and, in the case of independence, it would have the value zero. If the contrary evidence in the divisor exceeds the value in the numerator, it would be negative information, i.e., evidence against. The LDOR is popular for this and other reasons. Log odds convert logistic regression, especially when using a probability-based model, to a likelihood–based model that is easy to interpret. The LDOR also has a simple, meaningful, symmetric standard error SE. If the confidence interval CI is preferred, for a 95% CI, it is 1.96 times the SE. Although an effect of focusing on the disease is to increase the number of records, the relevant data may still frequently be sparse. The LDOR can also be estimated using *incompletely summated zeta functions* [28–31,39], which are measures of expected information. The approach has its roots in early bioinformatics [38].

$$\begin{aligned}\text{LDOR} \quad &= \zeta(s=1,\ n[\text{BPU},\ \text{conditions}]\ )\ +\ \zeta(s=1,\ n[\text{not(BPU)},\ \text{not(conditions)}]) \\ &-\zeta(s=1,\ n[\text{BPU},\ \text{not(conditions)}])\ )\ -\ \zeta(s=1,\ n[\text{not(BPU)},\ \text{conditions}])\end{aligned} \tag{4}$$

The above should be considered only as a recommendation because, in many cases, ζ(s=1, n) can be replaced by the natural logarithm ln(n). The importance of ζ( ) arises when information is sparse, though it must be recognized that such a situation is common. As one considers higher dimensional data, many joint events of interest have more attributes, and the number of their occurrences rapidly decreases. For present purposes, ζ(s, n) with s = 1 is simply the Euler series $1 + 1/2 + 1/3 + \ldots 1/n$, which suffices here. This equation is still referred to as the incomplete, partially summated, or generalized zeta function because ζ(s) assumes n = ∞ and other values of s have merit as surprise measures and moments of information, $\zeta\ (s, n)) = \Sigma_{i\ =\ 1,2,3\ \ldots\ n}\ i^{-s}$. Discussion of uses of the more general measure (with other values of s) is beyond the present scope. For extreme cases of sparse data when s = 1, ζ(s = 1, 0) = 0, ζ(s=1, 1) = 1, ζ(s=1, 2) = 1.5, and so on. This method avoids difficulties encountered in the traditional frequentist approach. It better reflects the *available information*, e.g., one observation out of 2 is not as informative as 100 out of 200. The exponential of the above LDOR, defined as information, is seen as the basis of a definition of DOR. However, if considered as an estimate, the traditional and above methods converge as follows.

$$\text{DOR} = \begin{matrix} e^{\zeta(s=1,\ n\ [\text{BPU},\ \text{conditions}]\ )\ -\ \zeta(s=1,\ n[\text{BPU}])\ -\ \zeta(s=1,\ n\ [\text{not(BPU)},\ \text{conditions}]\ )\ +\ \zeta(s=1,\ n[\text{not(BPU)}])} \\ \mathscr{L}\!\!\!/\ n[\ ] \rightarrow \infty \end{matrix} \tag{5}$$

When the target is a disease such as due to infection in a new epidemic, and there are several conditions, not(conditions) is usually a large number such that P(not(count)) ≈ 1, in which case P(disease, not (conditions)) ≈ P(disease) and P(not(disease), not(conditions)) ≈ P(not (disease). Then, the DOR and DOR* closely approximates the Likelihood Ratio LR by definition of the latter, and to a good approximation, the following apply.

$$\text{DOR*} \approx \text{LR} = \text{P(}\ \text{conditions}\ |\ \text{disease}\ )\ /\ \text{P(}\ \text{conditions}\ |\text{not(disease}\ )) \tag{6}$$

$$\begin{aligned}\text{LDOR*} \quad &= \zeta(s=1,\ n\ [\text{disease},\ \text{outcome good},\ \text{conditions}]) \\ &+\zeta(s=1,\ n[\text{not(disease)},\ \text{outcome good},\ \text{not(conditions)}]) \\ &-\zeta(s=1,\ n[\text{disease},\ \text{outcome good},\ \text{not(conditions}]) \\ &-\zeta(s=1,\ n[\text{not(disease)},\ \text{outcome good},\ \text{conditions}]) \\ &-\zeta\ (s=1,\ n\ [\text{disease},\ \text{outcome poor},\ \text{conditions}]\ ) \\ &-\zeta(s=1,\ n[\text{not(disease)},\ \text{outcome poor},\ \text{not(conditions)}]) \\ &+\zeta(s=1,\ n[\text{disease},\ \text{outcome poor},\ \text{not(conditions)}]\ ) \\ &+\zeta(s=1,\ n[\text{not(disease)},\ \text{outcome poor},\ \text{conditions}])\end{aligned} \tag{9}$$

Based on the previous reasoning, the DOR can usually be estimated by the LR when, as typically, the case P(not conditions) ≈ 1; however,

$$\text{LR}\ \approx e^{\zeta(s=1,\ n\ [\text{disease},\ \text{conditions}]\ )\ -\ \zeta(s=1,\ n[\text{disease})\ -\ \zeta(s=1,\ n\ [\text{not(disease)},\ \text{conditions}]\ )\ +\ \zeta(s=1,\ n[\text{not(disease)}])}. \tag{7}$$

one may use the following, usually to a good approximation, but note that it still involves eight probabilities.

$$\text{PO}\ \approx e^{\zeta(s=1,\ n\ [\text{disease},\ \text{conditions}]\ )\ -\ \zeta(s=1,\ n\ [\text{not(disease)},\ \text{conditions}]\ )} \tag{8}$$

If there are only two groups as in a placebo trial but, *with the important exception of the factor or factors that define any group*, even in a multifactorial 2 × 2 x 2 × … case, patients in any one group cannot be identical but rather are drawn at random regarding further factors. For an experimental RCT, this should (a) ideally be done in a manner that is intended to be representative of the larger population of future users of the therapy and in a manner respectful of strata in the population and equity, and (b) is also often said to be capable of determining whether a cause-effect relation exists between the intervention and the outcome. The above ideas are mutually contradictory and ambitious. They have more to do with the aspirations of the researchers rather than any analytical and predictive method. So, it is best to use bidirectional general graph methods, like those employed by Ingine, that do not presume which is cause, which is effect, and which have a common cause [34]. The first is by its nature not random, and the very notion of randomness is problematic for a small number of trial patients relative to the large number of future patients who may be prescribed the new therapy. Similar notions apply to purely computational approaches that are applied to drugs already approved and marketed.

As a follow-on from Eqn. (1) for the LR and Eqn. (3) for the DOR, and recalling that DOR* is a ratio of DORs, it is possible to see that the full and more accurate DOR* has eight probabilities. So, more precisely, we have the following.

$$\begin{aligned}\text{LDOR}^* &= \zeta\,(\text{s}=1,\ \text{n}\,[\text{disease},\ \text{outcome good},\ \text{conditions}]\,)\\ &+\zeta(\text{s}=1,\ \text{n}[\text{not(disease)},\ \text{outcome good})]\\ &-\zeta(\text{s}=1,\ \text{n}[\text{disease},\ \text{outcome good}\,])\\ &-\zeta(\text{s}=1,\ \text{n}[\text{not(disease)},\ \text{outcome good},\ \text{conditions}])\\ &-\zeta\,(\text{s}=1,\ \text{n}\,[\text{disease},\ \text{outcome poor},\ \text{conditions}]\,)\\ &-\zeta(\text{s}=1,\ \text{n}[\text{not(disease)},\ \text{outcome poor}\,])\\ &+\zeta(\text{s}=1,\ \text{n}[\text{disease},\ \text{outcome poor}\,])\\ &+\zeta(\text{s}=1,\ \text{n}[\text{not(disease)},\ \text{outcome poor},\ \text{conditions}])\end{aligned} \tag{10}$$

In trial design, insight is wanted into what further factors can be beneficial or otherwise for a patient; it is both natural and responsible to ask why and how a particular single patient responded adversely to a drug during a trial, and hence to try and avoid such a situation from the outset. Unfortunately, it can be argued that injudicious multiple subgroup analyses of the same data increase the risk of generating false positive findings. All outcomes and planned subgroup analyses should thus be pre-specified and described in the original trial registry. But clearly, retrospective cohort studies over large numbers of records can help the latter aim. A great deal of what is considered as randomization is really an expression of unjustified ignorance since a detailed retrospective cohort study could always "find out," data permitting, and few things are truly independent. A better notion, like the one employed by Ingine, is not that things are randomly associated but that they are treated as "doesn't matter," "not taken into account," or more simply as combined or pooled. For example, if 'something':='Y' or 'something':='N' are specified on the list of interdependent and independent factors, then the intention is clear. At the same time, if neither is mentioned, it means that the rest will apply whether 'something':='Y' or 'something':='N' occurs. In the interpretation of being "pooled," the relevant prevalences implied, e.g., P('something':='Y') and P('something':='N'), reflect the population in the records analyzed, which may or may not be extensible (the former being the normal assumption of statistics). For example, a consequence is if all records contained 'something':='Y' or unknowns, and none contain 'something':='N'. Neither specifying 'something':='Y' nor 'something':='N' is the same as specifying 'something':='Y.'

3. Methods

Although no medical record data as yet exists for a future pandemic of the likely severity of avian influenza, the proprietary historical data set used here for testing and comparing methods consists of structured patient records [33] including standard extensive demographic and comorbidity data along with bloodwork tests collected with medical personnel in the field [28], a large set of CMS healthcare insurance claim records [33], and socioeconomic health data that allowed some joining by US County or State in a few experimental runs. They have been tested in a variety of studies including many cited here. A variety of automated methods have been implemented to test the quality of data and ensure correct management of unknowns [39] which arise extensively in clinical lab results and joined data. In addition, as well as our own data, knowledge gathering methods [29,30,31 including those used in the early COVID-19 studies [10,12–15] contributed to the tables presented in Results Section 4.1. These represent data already mined or otherwise analyzed by other workers, which could be collated and converted to the measures of interest here (including novel measures such as DOR* and Severity Score described below).

It is appropriate to mention two recent major AI developments which should be applicable. Rohekar et al. argue in favor of Deep Learning that makes use of or learns from Bayes Nets [78], but this argument has been challenged [79]. Although the inclusion of a Bayes Net component has the advantage of moving the technology toward a neural net with a more explanatory ontological structure discussed in Theory Section 2.1, ensembles of DL networks better explore the loss or fit function surface without being so easily trapped in a local solution [79]. The relative merits may, however, may well depend on (i) how many exploration trajectories are used in an ensemble approach and (ii) whether the Bayesian structure information used or acquired is a good or bad representation of the problem being studied. From the latter perspective, a full glass box with step-by-step visible use of the data from the house may be a better choice. For example, as discussed in Results Section 4, the query entries used in the construction of large glass box inference nets by partially supervised data mining [33,39] can be populated by attributes associated with the prediction target as found by unsupervised data mining. Here, "unsupervised" means that specific target attributes or clusters of interest are not pre-stipulated. However, all data may be labeled, e.g., by using attribute type names as column headers. The second development is the rise of the whole "explainable AI (XAI) community XAI (which has some overlap with the above). The prediction strategies of these models at times turned out to be frequently flawed and not aligned with human intuition, probably due to false associations due to fluctuation in sparse data, effectively the kind of problem that the zeta function approach seeks to solve [33,38,39] and as shown in Eqns. 4-9 above. Current XAI is focused on solving the above difficulty and improving model efficiency, robustness, and generalization [80], though not to our knowledge in a way like the zeta approach. However, several XI researchers have seen a solution in *feature selection* or *feature extraction* [81], which was discussed at the end of Introduction Section 1.2, in Theory Section 2.4, and relation to the overview by Shah et al. [59] concerning DL and clinical trials. It is also essentially the use of something like unsupervised data mining in relation to Bayesian *versus* DL ensembles immediately above.

As stated by Landolsi and colleagues in a comprehensive review of information extraction from electronic medical documents [82], the main feature of information extraction is data mining. They describe the traditional approach is comprising five steps: (i) data collection, (ii) selection of target data, (iii) preprocessing and transformation of the data (which many authors would describe as curation), (iv) identification of patterns by data mining (which many authors would describe as clusters although negative associations are as important as positive associations), and (v) interpretation and evaluation as knowledge extraction. These might be compared with the notions of self-organizing maps (e.g. Ref [83]) and random forest methods (e.g. Ref [84) that are often considered black box but have relationships with patterns and clusters. Two observations by the above reviewers [82] are of interest here. First, information extraction is seen as a step within text mining that is like pre-processing in the classical data processing procedure. Second, those reviwers consider hand-coded rules to give a better result than ML methods, although requiring manual construction of rules with the assistance of a medical expert. What would the results of this process look like as visible, explanatory, probabilistic "rules" gathered by more automatic methods and intended to be used in predictive mode? Refs [33,38,39] describe methods that can also perform unsupervised structured data mining to obtain the kind of tags below, and similar tags can be ultimately generated by auto-surfing the internet for relevant authoritative text [29,30]. Two examples are as follows.

<Q-UEL-COVID19-BY-COUNTRY 'variant':='eqB.1.1.7′ Pfwd:=0.8896 Ofwd:=3.9606 Efwd:=0.6371 | **if**:=(assoc:=3.4616, count:=7, factors:=(2,5)) | 'percent variant':='ge50' 'percent cases sequenced':='0.2' 'valid denominator':='Yes' **with** 'percent cases sequenced':='0.2' 'source':='TESSy' 'country':='Romania' Pbwd:=0.006057 Obwd:=11.4507 Ebwd:=1.8420 Q-UEL-COVID19-BY-COUNTRY>

<Q-UEL-CTRACT 'Severe disease':='Y' Pfwd:=0.108 | **if** | Infection:=COVID-19 'Based on':=First wave death':='Y' 'type 2 diabetes':='Y' 'prior cerebrovascular disease':='Y' Pbwd:=0.021 Q-UEL-CTRACT>

These are some of the simpler examples using the Q-UEL language, an extension of XML for probabilistic semantics based on the Dirac notation and algebra, as used in coauthor BR's approach [21–39] first discussed in Section 1.2. The important point is that the Diracbraket $<A|B>$ written here in format $<A|$ **if** $|B>$ encodes the probability dual $\{P(A|B), P(B|A)\}$ as a particular kind of complex number which is not

necessary to discuss for present purposes: it is a particular glass box approach and the general idea should be intuitive as follows. Each of *A* and *B* here can represent a variety of attributes using the ':=' notation discussed above, and each is more generally an expression, typically a logical expression, containing them. Here, it suffices to consider all attributes in the tag as linked by logical AND (except for the logical IF). In the above Q-UEL tag, Pfwd is the probability forward for $<A|B>$ as P(A|B), and Pbwd is its adjoint, the probability backward P(B|A). In the more explicit medical notation of Theory Section 2.5, the Pfwd value of form P (disease | conditions) and the Pwd value is of form P(conditions | disease). While other features are of less interest here, it is also worth noting that Obwd ("odds backward") means the Likelihood Ratio LR, which under these conditions is a close approximation of the DOR.

The kind of information that should be made available from the analysis of many patient records can generate the above Q-UEL XML-like"tags." Many such tags can automatically build and form the building blocks of an odds inference net for predictions of LR and DOR or analogous measures for diagnosis, risk assessment, or choice of best therapy [33,39]. However, black-box and glass-box methods, including inference nets, are really estimates that make implicit or explicit independence assumptions when there is inadequate data. The above glass-box methods used are such algorithms as used by Ingine, including the unusual property that the inference nets will automatically adjust when there is sufficient data for an exact calculation. This approach is in the manner of those Evidence-Based Medicine or epidemiological measures that are frequently based on an exact count in the manner of a census (though inevitably, they are seen as based on a restricted sample when future use is made of them). The algorithms then deduce the remaining interactions by data mining that extends the number of demographic and clinical factors, building on that basis by a perturbation method [33,39]. This approach provides a means of including information in the present paper at the outset when probabilistic measurements are available from epidemiological studies. For example, the compilation of statistics from Barron et al. [66] was used extensively in the study but extensively supplemented by means of capturing knowledge from the Internet and converting it to Q-UEL tags [10–14,19,29,30].

The table generated In Results Section 4 represents predictions or estimates of risks of *severe COVID* in terms of roughly estimated measures from a large range of literature and government releases related to the first wave of COVID-19, especially in the UK. The measures used are association constants of general form P(A, B)/P(A)P(B), Predictive Odds P(A|B)/P(not-A|B), Likelihood Ratio P(B|A)/P(B|not-A), where A is *severe COVID* and B represents a set of various conditions such as demographic and clinical factors. In the specific infectious disease ontology of Theory Section 2.2, A is more precisely defined as patient state C followed by R (complications with recovery) or with patient state C followed by D (complications followed by death), as C, R, and D are defined above in Theory Section 3.1. More important than the value estimates for the above measures is that they list the demographic and clinical factors that need to be inspected in any black-box or glass-box machine learning in the case of new and emerging epidemics. However, those presented here are by no means exclusive since new symptoms and physiological considerations may well apply. The actual values quoted give this study much of the flavor of a systematic review and meta-analysis, though that is not the intent. However, the numbers represent the kind of values that might be expected for an emerging pandemic, and the relative values or ranking of conditions will be more plausible than the absolute values.

The fact that COVID and avian influenza are extensive respiratory diseases and the possibility that both might have originated as alimentary tract infections means that several similarities might well be expected to arise in the case of avian influenza. A novel measure used by Ingine called the *Severity Score* SS is introduced below. Expressed, this relates to the prevalence of a particular set of features of a disease that suggest high risk for specified kinds of patients, and hence the fraction of

Table 1
Example estimates of risks for severe respiratory viruses for diabetic and non-diabetic patients conditional on standard demographic factors based on fatalities in the first wave of COVID-19.

Probability P( ) of severe COVID-19-like disease is considered as the probability of state C (complications with recovery) or with state (D) exclusive, conditional upon the following factors.	Association Constant	Predictive Odds assuming high 5% prevalence of severe disease.	Likelihood Ratio	SS = SEVERITY SCORE Info = 10-ln(P) to nearest integer
First wave, COVID-19, no diabetes, Age 0-39	0.1	0.005	0.1	2
First wave, COVID-19, no diabetes, Age 40-49	0.2	0.011	0.2	3
First wave, COVID-19, no diabetes, Age 50-59	0.6	0.032	0.6	4
First wave, COVID-19, no diabetes, Age 60-69	1.8	0.095	1.8	5
First wave, COVID-19, no diabetes, Age 70-79	4.9	0.263	5.0	6
First wave, COVID-19, no diabetes, Age 80+	20.2	1.153	21.9	7
First wave, COVID-19, type 1 diabetes, Age 0-39	1.4	0.074	1.4	5
First wave, COVID-19, type 1 diabetes, Age 40-49	2.8	0.153	2.9	5
First wave, COVID-19, type 1 diabetes, Age 50-59	5.7	0.305	5.8	6
First wave, COVID-19, type 1 diabetes, Age 60-69	11.5	0.632	12.0	7
First wave, COVID-19, type 1 diabetes, Age 70-79	22.8	1.316	25.0	8
First wave, COVID-19, type 1 diabetes, Age 80+	59.6	4.053	77.0	9
First wave, COVID-19, type 2 diabetes, Age 0-39	1.5	0.079	1.5	5
First wave, COVID-19, type 2 diabetes, Age 40-49	2.4	0.126	2.4	5
First wave, COVID-19, type 2 diabetes, Age 50-59	4.4	0.237	4.5	6

(*continued on next page*)

Table 1 (*continued*)

Probability P( ) of severe COVID-19-like disease is considered as the probability of state C (complications with recovery) or with state (D) exclusive, conditional upon the following factors.	Association Constant	Predictive Odds assuming high 5% prevalence of severe disease.	Likelihood Ratio	SS = SEVERITY SCORE Info = 10-ln(P) to nearest integer
First wave, COVID-19, type 2 diabetes, Age 60-69	8.2	0.442	8.4	7
First wave, COVID-19, type 2 diabetes, Age 70-79	15.5	0.868	16.5	7
First wave, COVID-19, type 2 diabetes, Age 80+	37.6	2.305	43.8	8
First wave, COVID-19, other diabetes, Age 0-39	0.6	0.032	0.6	4
First wave, COVID-19, other diabetes, Age 40-49	1.1	0.058	1.1	5
First wave, COVID-19, other diabetes, Age 50-59	2.3	0.126	2.4	5
First wave, COVID-19, other diabetes, Age 60-69	4.7	0.247	4.7	6
First wave, COVID-19, other diabetes, Age 70-79	17.3	0.979	18.6	7
First wave, COVID-19, other diabetes, Age 80+	35.9	2.189	41.6	8
First wave, COVID-19, type 1 diabetes, male	35.3	2.142	40.7	8
First wave, COVID-19, type 1 diabetes, female	26.2	1.532	29.1	8
First wave, COVID-19, type 2 diabetes, male	68.3	4.853	92.2	9
First wave, COVID-19, type 2 diabetes, female	47.4	3.037	57.7	8
First wave, COVID-19, other diabetes, male	46.2	2.953	56.1	8
First wave, COVID-19, other diabetes, female	27.3	1.605	30.5	8
First wave, COVID-19, no diabetes, male	7.5	0.405	7.7	6
First wave, COVID-19, no diabetes, female	5.0	0.268	5.1	6
First wave, COVID-19, no diabetes, Asian	1.7	0.089	1.7	5
First wave, COVID-19, no diabetes, Black	2.4	0.126	2.4	5
First wave, COVID-19, no diabetes, Mixed	0.7	0.037	0.7	4
First wave, COVID-19, no diabetes, Other	1.3	0.068	1.3	5
First wave, COVID-19, no diabetes, White	2.3	0.121	2.3	5
First wave, COVID-19, no diabetes, Unknown	0.6	0.032	0.6	4
First wave, COVID-19, type 1 diabetes, Asian	20.8	1.189	22.6	7
First wave, COVID-19, type 1 diabetes, Black	28.7	1.695	32.2	8
First wave, COVID-19, type 1 diabetes, Mixed	9.0	0.489	9.3	7
First wave, COVID-19, type 1 diabetes, Other)	14.9	0.832	15.8	7
First wave, COVID-19, type 1 diabetes, White)	7.9	0.426	8.1	6
First wave, COVID-19, type 1 diabetes, Unknown	2.1	0.111	2.1	5
First wave, COVID-19, type 2 diabetes, Asian	15.1	0.842	16.0	7
First wave, COVID-19, type 2 diabetes, Black	37.5	2.300	43.7	8
First wave, COVID-19, type 2 diabetes, Mixed	22.4	1.289	24.5	8
First wave, COVID-19, type 2 diabetes, Other	18.1	1.021	19.4	7
First wave, COVID-19, type 2 diabetes, White	17.3	0.974	18.5	7

(*continued on next page*)

Table 1 (*continued*)

Probability P( ) of severe COVID-19-like disease is considered as the probability of state C (complications with recovery) or with state (D) exclusive, conditional upon the following factors.	Association Constant	Predictive Odds assuming high 5% prevalence of severe disease.	Likelihood Ratio	SS = SEVERITY SCORE Info = 10-ln(P) to nearest integer
First wave, COVID-19, type 2 diabetes, Unknown	4.6	0.247	4.7	6
First wave, COVID-19, other diabetes, Asian	9.6	0.526	10.0	7
First wave, COVID-19, other diabetes, Black	15.9	0.889	16.9	7
First wave, COVID-19, other diabetes, Mixed	4.9	0.263	5.0	6
First wave, COVID-19, other diabetes, Other	9.6	0.526	10.0	7
First wave, COVID-19, other diabetes, White	12.4	0.684	13.0	7
First wave, COVID-19, other diabetes, Unknown	3.3	0.179	3.4	6

the population that will need intensive hospital care. The notion of "severity" is intended to apply to the severity of the epidemic as far as particular strata of the community are concerned, e.g., an ethnic or age group or patients with type 2 diabetes. Below, the score SS is the nearest integer to 10 – lnP((C & R) or (C & D)), with P( ) originally from the first-wave COVID-19 analysis. However, it is intended that this is adjusted so that a score spanning 0, 1, 2, …10 is obtained for use in future respiratory virus pandemics, as discussed in Results Section 4.1. Relevant historical studies that provide a tangible example may be reported here, which this paper seeks to generalize. The first is based primarily on a compilation of studies, especially that by Barron et al. [66], for the first wave of COVID-19 in 2020 but contains data that was obtained from patient data of the kind that can or should be found on rapidly updated digital patient records. Based on that compilation, the following are two examples from Ref. [19] of knowledge elements expressed in Ingine's Q-UEL language, generated and/or used in prediction by Ingine modules such as DiracMiner or DiracSmash [33,39].

4. Results

4.1. Examples of estimates of risk dependent on different demographic and clinical factors

An important aim for generating these results is to highlight the attributes of patients that focus on early analysis of a future respiratory virus epidemic, particularly (but not necessarily only [59,74]) with the use of large numbers of regularly updated digital patient records. In that sense, they could form a preliminary "prepackaged" feature selection

Table 2
Example estimates of risks for severe respiratory viruses for diabetic and non-diabetic patients conditional on deprivation class and comorbidities based on fatalities in the first wave of COVID-19.

Probability P( ) of severe COVID-19-like disease is considered as the probability of state C (complications with recovery) or with state (D) exclusive, conditional upon the following factors.	Association Constant	Predictive Odds assuming high 5% prevalence of severe disease.	Likelihood Ratio	SS = SEVERITY SCORE Info = 10-ln(P) to nearest integer
First wave, COVID-19, no diabetes, Deprivation rank:=0	2.2	0.116	2.2	5
First wave, COVID-19, no diabetes, Deprivation rank:=1	1.6	0.084	1.6	5
First wave, COVID-19, no diabetes, Deprivation rank:=2	1.5	0.079	1.5	5
First wave, COVID-19, no diabetes, Deprivation rank:=3	1.5	0.079	1.5	5
First wave, COVID-19, no diabetes, Deprivation rank:=4	1.4	0.074	1.4	5
First wave, COVID-19, no diabetes, Deprivation rank:=5	1.4	0.074	1.4	5
First wave, COVID-19, no diabetes, Deprivation rank:=unknown	1.5	0.079	1.5	5
First wave, COVID-19, type 1 diabetes, Deprivation rank:=0	18.3	1.037	19.7	7
First wave, COVID-19, type 1 diabetes, Deprivation rank:=1	10.3	0.563	10.7	7
First wave, COVID-19, type 1 diabetes, Deprivation rank:=2	8.9	0.484	9.2	7
First wave, COVID-19, type 1 diabetes, Deprivation rank:=3	8.0	0.432	8.2	7
First wave, COVID-19, type 1 diabetes, Deprivation rank:=4	5.3	0.289	5.5	6
First wave, COVID-19, type 1 diabetes, Deprivation rank:=5	4.2	0.226	4.3	6
First wave, COVID-19, type 1 diabetes,	7.3	0.400	7.6	6

(*continued on next page*)

Table 2 (*continued*)

Probability P( ) of severe COVID-19-like disease is considered as the probability of state C (complications with recovery) or with state (D) exclusive, conditional upon the following factors.	Association Constant	Predictive Odds assuming high 5% prevalence of severe disease.	Likelihood Ratio	SS = SEVERITY SCORE Info = 10-ln(P) to nearest integer
Deprivation rank:=unknown				
First wave, COVID-19, type 2 diabetes, Deprivation rank:=0	28.4	1.674	31.8	8
First wave, COVID-19, type 2 diabetes, Deprivation rank:=1	16.0	0.895	17.0	7
First wave, COVID-19, type 2 diabetes, Deprivation rank:=2	15.7	0.879	16.7	7
First wave, COVID-19, type 2 diabetes, Deprivation rank:=3	13.2	0.732	13.9	7
First wave, COVID-19, type 2 diabetes, Deprivation rank:=4	12.2	0.668	12.7	7
First wave, COVID-19, type 2 diabetes, Deprivation rank:=5	11.6	0.642	12.2	7
First wave, COVID-19, type 2 diabetes, Deprivation rank:=unknown	13.7	0.763	14.5	7
First wave, COVID-19, other diabetes, Deprivation rank:=0	24.0	1.395	26.5	8
First wave, COVID-19, other diabetes, Deprivation rank:=1	12.1	0.668	12.7	7
First wave, COVID-19, other diabetes, Deprivation rank:=2	9.7	0.532	10.1	7
First wave, COVID-19, other diabetes, Deprivation rank:=3	9.0	0.489	9.3	7
First wave, COVID-19, other diabetes, Deprivation rank:=4	7.7	0.416	7.9	6
First wave, COVID-19, other diabetes, Deprivation rank:=5	7.3	0.395	7.5	6
First wave, COVID-19, other diabetes, Deprivation rank:=unknown	9.1	0.500	9.5	7
First wave, COVID-19, type 1 diabetes, prior coronary heart disease	20.2	1.147	21.8	7
First wave, COVID-19, type 2 diabetes, prior coronary heart disease	19.0	1.079	20.5	7
First wave, COVID-19, other diabetes, prior coronary heart disease	15.2	0.853	16.2	7
First wave, COVID-19, no diabetes, prior coronary heart disease	7.6	0.411	7.8	6
First wave, COVID-19, type 1 diabetes, prior cerebrovascular disease	32.0	1.921	36.5	8
First wave, COVID-19, type 2 diabetes, prior cerebrovascular disease	28.3	1.674	31.8	8
First wave, COVID-19, other diabetes, prior cerebrovascular disease	24.5	1.421	27.0	8
First wave, COVID-19, no diabetes, prior cerebrovascular disease	11.5	0.632	12.0	7
First wave, COVID-19, type 1 diabetes, prior heart failure	35.9	2.184	41.5	8
First wave, COVID-19, type 2 diabetes, prior heart failure	30.7	1.832	34.8	8
First wave, COVID-19, other diabetes, prior heart failure	23.4	1.353	25.7	8
First wave, COVID-19, no diabetes, prior heart failure	14.9	0.826	15.7	7

Table 3
Examples of estimated risks from early symptoms associated with later severe respiratory virus infection in patients at high risk due to comorbidities.

Probability P( ) of severe COVID-19-like disease is considered as the probability of state C (complications with recovery) or with state (D) exclusive, conditional upon initial COVID diagnosis with at least two of common comorbidities (type 1 or type 2 diabetes, chronic kidney disease, coronary heart disease, cerebrovascular disease, or heart failure), and the following initial symptoms.	Association Constant	Predictive Odds assuming high 5% prevalence of severe disease.	Likelihood Ratio	SS = SEVERITY SCORE Info = 10-ln(P) to nearest integer
Severe headache	222.9	76.674	1456.8	10
Pain in the pharynx	222.6	76.089	1445.7	10
Runny nose	218.4	30.958	588.2	10
Respiratory distress, shortness of breath, continuous cough or wheezing	202.9	46.632	886.0	10
Sinus headache	181.6	30.826	585.7	10
Sore throat	181.6	30.826	585.7	10
Sneezing	178.9	28.958	550.2	10
Fatigue	166.8	33.911	644.3	9
Muscle ache	158.2	20.863	396.4	9
Diarrhea	123.7	12.284	233.4	9
Expectoration	105.8	9.311	176.9	9
Treatment by drugs prior to admission	91.1	7.326	139.2	9
Chest pain or tightness	88.2	6.979	132.6	9
Loss of taste and smell	86.6	6.789	129.0	9
Weakness	66.8	4.716	89.6	9
New continuous cough	48.4	3.121	59.3	8
Fever high or shivering (chills)	10.0	0.547	10.4	7
Slightly raised temperature	8.9	0.489	9.3	7

[59,81], discussed at several points above, for both glass box and black box methods. More formally and automatically, they could be used as Bayesian priors, as discussed in the next paragraph below. Although, as indicated above, these results and other results in the Sections below were (a) obtained with the help of tools being commercialized by Ingine and (b) are intended to illustrate the use of the measures recommended by the present authors, they are of a nature that could be reproduced by generally available glass box methods [82] or simple adaptations of them. That is, it will also aid in feature selection and "explainability" for ML and DL approaches. There is one important exception to the use of more widely used tools, and that is the involvement of the partially summated zeta function discussed in Section 2.6 for the treatment of sparse data, which has its roots in early bioinformatics methods [38]. The measures in Tables 1–4 (not previously published) but derived and

Table 4
Examples of estimated risks from clinical test results associated with later severe respiratory virus infection in patients for patients at high risk due to comorbidities.

Probability P( ) of severe COVID-19-like disease is considered as the probability of state C (complications with recovery) or with state (D) exclusive, conditional upon initial COVID diagnosis with at least two of common comorbidities (type 1 or type 2 diabetes, chronic kidney disease, coronary heart disease, cerebrovascular disease, or heart failure), and the following early clinical factors.	Association Constant	Predictive Odds assuming high 5% prevalence of severe disease.	Likelihood Ratio	SS = SEVERITY SCORE Info = 10-ln(P) to nearest integer
30 or more breaths/minute, a blood oxygen saturation less than 94% or less, and a ratio of the partial pressure of arterial oxygen to the fraction of inspired oxygen less than 300 mm Hg	252.6	332.421	6316	10
Extensive lung opacity in X-ray	252.0	312.789	5943	10
Neutrophil/ lymphocytes ratio greater than 5.86	251.6	300.947	5718	10
Indirect bilirubin high	239.5	140.053	2661	10
CK-MB high	213.4	59.421	1129	10
Alanine transaminase high	203.4	47.163	896.1	10
Prothrombin International Normalized Ratio high	153.9	19.526	371.0	9
Blood urea nitrogen high	147.4	17.626	334.9	9
Aspartate transaminase high	146.6	17.416	330.9	9
Total bilirubin high	144.2	16.789	319.0	9
CT Bilateral brain lesion	102.1	8.779	166.8	9
D-dimer high	86.1	6.732	127.9	9
Creatinine high	84.2	6.516	123.8	9
Aldehyde dehydrogenase high	53.0	2.300	43.7	9
Prothrombin time high	57.6	3.884	73.8	8
Red blood cells low	55.5	3.705	70.4	8
Direct bilirubin high	52.6	3.463	65.8	8

(*continued on next page*)

Table 4 (*continued*)

Probability P( ) of severe COVID-19-like disease is considered as the probability of state C (complications with recovery) or with state (D) exclusive, conditional upon initial COVID diagnosis with at least two of common comorbidities (type 1 or type 2 diabetes, chronic kidney disease, coronary heart disease, cerebrovascular disease, or heart failure), and the following early clinical factors.	Association Constant	Predictive Odds assuming high 5% prevalence of severe disease.	Likelihood Ratio	SS = SEVERITY SCORE Info = 10-ln(P) to nearest integer
White blood cells abnormal	32.4	1.942	36.9	8
Creatine kinase high	31.8	1.905	36.2	8
Albumin low	11.8	0.653	12.4	7
Platelets low	7.6	0.416	7.9	6
Neutrophils high	4.5	0.242	4.6	6
C-reactive protein high	2.4	0.126	2.4	5

collated from diverse sources by knowledge gathering tools (Methods Section 4) similar to those used by coauthor BR in the early response to COVID-19, are intended to be useful for a future respiratory virus epidemic but are deduced from the first wave of COVID-19 in the UK from 23rd March to May 30, 2020. The statistics in the tables overall were first collated by author BR in mid-October 2020 during the early phase of the second wave on 7th September. Tables 1 and 2 rest heavily on the work of Barron and colleagues [66]. Still, the metrics used are different, adapted to represent severe disease from new respiratory virus pandemics in general and in different conditions. To a first rough approximation, it was reasonable to consider the case fatality rate in the first wave as translating to severe COVID-19 in the second and later waves (albeit with a risk of death). That is because it was becoming apparent that the case rate of fatalities was hugely reduced in the second wave. In hindsight, this was by about 83.3%. The second and later waves were not analyzed in as much detail, not least because, in the present paper, we are primarily interested in the challenges provided by the first waves of a new pandemic, notably a dangerous, radically different new strain of SARS-Cov-2 or Avian Influenza in humans.

While our collaborators at ETRI focus on developing ML and DL methods, equations like Eqn. (4) based on the zeta function ζ(s, n), used by Ingine as the basic elements of a glass-box method that can manage sparse as well as plentiful data, and they also provide a powerful and natural way to make use of previous pandemics. This approach can be shown to arise from integrals in which ζ(s, n+v) is equally valid, where v is a "virtual frequency" [28,33,39]. This can be easily derived from previous studies (such as on earlier pandemics) as well as Bayesian prior degrees of belief, e.g., v = P × N, where P is a prior probability and N is the total amount of data available. Note that as n increases, a smaller contribution from prior or subjective information represented through v is naturally weakened in impact. Also, keeping ζ(s, n) for currently observed occurrences but including Q-UEL tags based on ζ(s, v) is a convenient and slightly different estimate that nonetheless still gives results that converge rapidly and to traditional estimates as n and v increase. While it could be said that tags based on ζ(s, v) were used in that spirit to capture information from the sources used for Tables 1–4, it is perhaps more appropriate to say that ζ(s, n) was used. This is because epidemiological source data such as in Ref. [66] almost always reports counts n, which can be used directly or reliably inferred from the information given. Also, a modified zeta function [28] (originally intended to be applied to uses of a single zeta function) was found to give very similar results in the present study due to one zeta function always being subtracted from an appropriate other. In this phase of the study, Ingine's DiracSmash, or more precisely a variant form called QuickPredict, works with tags computed from zeta functions collated from very diverse sources [39] and can be used to combine information from several sources [33,39]. However, it is also a convenient way, with multiple cross-checks on consistencies between probabilities and measures, to compute probabilities, Predictive Odds PO, and Likelihood Ratios LR from which other metrics such as association constants as DOR* can be calculated where appropriate, even when directly based on counts n as above. Note that probabilities, PO, LR, and association constants are also available on the individual tags, as shown in Methods Section 3. In contrast, the Severity Scores in the last column of each table were computed directly from the source data as described below.

In this first part of the study, Q-UEL tags derived from the mining of patient records were not included. However, Q-UEL tags generated by searching the Internet for historical data were employed. However, see Section 4.2 (the case of including tags from digital patient records implies the use of synthetic data, which is necessary because no directly relevant records for a future pandemic are available). Consequently, it is also possible to verify results against those computed more directly (and more laboriously) from the original source data, providing the following generalizations are also considered, which make Tables 1–4 appropriate for future pandemics. Ingine's DiracSmash and QuickPredict [39] were convenient because they also involve statistical measures that are less likely to change for a new pandemic, i.e., giving input from experience that is better transferred from prior to new pandemics is, of course, important. Association constant and LR are given in Tables 1–4 because they are much less sensitive to prevalences and are more likely to apply (say as representing a prior degree of belief) to deductions about a new pandemic based on an old one such as the first wave COVID-19. Note that association constants and LRs have similar values if the probability for conditions P(conditions) is very small. Consequently, P(severe disease, not-conditions) is close to P(severe disease). This is often the case.

In contrast, the prevalence, Prior Odds PrO, and Predictive Odds PO are the most variable factors for a different pandemic or even a new wave of a similar one. Recall that the second wave of COVID was about three times the size of the first in many nations. The Prior Odds PrO is the prevalence of disease not conditional on factors, in this case, the prevalence of severe disease due to a pandemic. The Predictive odds PO are indicative of the prevalence of a severe form of a disease for a specified group, i.e., given a set of conditions such as demographic and clinical factors. When PrO is needed, e.g., for calculating the number of hospital beds required for a nation, the simplest way to compute the prevalence of the severe disease (as a percentage) is to take whatever the PO value is currently considered to be (e.g., note Eqn. (12) below) and use the following where LR is again the Likelihood Ratio.

$$\text{prevalence}\% = 100\ \text{PrO} \,/\, (1 + \text{PrO}) = 100\ \text{PO} \,/\, (\text{PO} + \text{LR}) \tag{11}$$

Note that when the Prior Odds PrO is available, LR is not required.

Conversely, responders almost certainly wish to specify a prevalence for a new pandemic to set sensible relationships between the above and other measures. However, that requires that Tables 1–4 use a standard reference PO, which gives a reasonable guess for a typical pandemic but can be adjusted by the actual prevalence for serious forms of the disease, say s. Since we are modeling potentially serious future strains but with reasonably plausible prevalences, we here assume 5% = 100s for a potential future serious disease in the sense comparable with serious

COVID-19 disease, i.e., corresponding beginning to patients states C followed by R and C followed by D as defined at the beginning of Methods Section 3.1. This choice of 5% needs some explanation. Recall that the Omicron BA.1 variant peaked in London at 8.79% of the population infected and, in the UK, overall at 2.69% on January 7th' 2022. In global estimates, 80% of infected people with COVID-19 suffered a mild or moderate form of the disease, 15% required hospitalization, and around 5 percent were critically ill.

By the considerations of Introduction Section 1.5, 1%, and almost certainly 2–5%, the newly seriously ill population could overwhelm most healthcare systems: the UK has *critical* beds for 0.0073%, and most other industrial nations are very similar, except Germany and the US with about 0.034%. But that is by no means the most disastrous pandemic imaginable. Admittedly, in the COVID-19 peak in the UK in January 2021, the fraction of the population hospitalized in standard and critical beds combined was about 0.06%, so 2–5% would appear excessive. Nonetheless, it is not a worst-case scenario. Recall that an estimated 33% of the world's population was infected and had clinically obvious symptoms during the 1918–1919 influenza pandemic. Influenza case-fatality rates were greater than 2.5%, compared to less than 0.1% in other influenza pandemics. Worse still, such numbers are dwarfed historically by the Black Death and by future potential widespread avian influenza or hemorrhagic fever epidemics in humans even armed with modern medical technology. On these grounds, 5% appears extremely optimistic.

Moreover, even given the number of critical beds in the US, 20% or more would seem more appropriate in view of the percentage levels of even just susceptible patients because of current chronic lung disease in the US (Section 4.2). So, to be cautious, it is better stated that the numbers to be considered in the Predictive Odds column should be multiplied by s(0.95)/(0.05(1-s)) where s is the prevalence of serious disease in the population as a fraction (i.e., on the interval 0 … 1). The new PO as PO*, where PO is the number *currently* specified in each table, is then as follows.

$$PO^* = 19 \cdot PO \cdot s \,/\, (1 - s) \tag{12}$$

Inserting Eqn. (11) to base it on the PO prior to adjustment is, of course, straightforward but leads to a less memorable equation.

In contrast to the above, Severity Scores given in the last columns of the Tables were always computed independently from source data. Original raw estimates (see below) were generally reduced in the second and subsequent COVID-19 waves, plausibly because the duration from symptoms to hospital admission was lower during the second wave, diagnostic testing was more widespread in the second wave, and on December 2, 2020, the Pfizer-BioNTech vaccine was approved for use in the UK, followed on 30 December by the cheaper and easier-to-distribute Oxford-AstraZeneca vaccine. As calculated here, epidemics are based on the probabilities of severe COVID deaths conditional upon various factors through the empirical measure 10 – ln(P), here interpreted as 10 – lnP((C & R) or (C & D)) (recall, C = complications, R = recovery, D = death) because fatalities in first wave COVID-19 are interpreted as later analogous to complication with recoveries as well as fatalities as discussed above. The remaining high scores in the second and subsequent waves are primarily associated with patients with comorbidities such as diabetes, and this is therefore emphasized in Tables 1–4 All these together represent the original "raw estimates" mentioned above. It is recommended that a similar scale of 0–10 be used to retain the spirit of the usefulness of these tables for future pandemics based on scaling P appropriately, e.g., so that patients less than 40 years old and without any comorbidities, symptoms, and abnormal clinical lab results, score 2, while patients aged 80 or more with at least one comorbidity score 8, leaving 0, 1 for those who are reliably vaccinated or not exposed, and 9, 10 for patients are already at extreme risk, say with chemotherapy etc. On that basis, it is also reasonable to say that the values in the last columns of Tables 1–4 also represent the final generalized SS measures.

4.2. Comparison of glass-box and black box methods

AI techniques of the black-box ML and DL type will be powerful aids, and measures like those above can be computed by assessing the quality of predictions of the clinical and other consequences of a new pandemic, but some requirements limit applicability. There must be a reasonably high degree of analogy between past epidemics and a new one and enough experience with a new pandemic, which takes time, to ensure that the experience from history is relevant. Focusing only on serious respiratory diseases, which often have analogies, would help (see Section 4.3 below), but glass-box methods should always be applicable. This is because they make visible both new and historically relevant conditions and probabilities needed for adaptability and extensibility to future cases. In this paper, the above results apply to cases modeled on the first COVID-19 wave for which exact calculations were possible and extensible by further data mining.

Though extensive data is not yet available for a very severe respiratory virus pandemic with fatality rates at the level that are expected for avian influenza, two approaches are possible. One is to construct synthetic data in a way that retains realism (Section 4.3 below), and the other is simply to assess comparative of different methods using past patient medical records to predict other, more familiar, diseases. Ref [60] gives some indication of the general performance of black box machine learning in comparison to the glass box methods used here. The glass box approach used was primarily the DiracSmash odds inference nets [33,39], with results that may be summarized briefly as follows because they give some interesting insights. Unsupervised data mining normally used to construct inference nets [28,34] also played a role by helping select suitable demographic and clinical descriptions of model patients characteristic of having, or not having, the disease. For three different sets of (kinds of) patients represented by three different sets of demographic and clinical descriptions of patients, predictions were made of congestive heart failure using structured data of 198 attributes and variously 500,000 to 700,000 records (depending on how the data analogous to training, i.e. essentially the data datamined, and testing sets were separated). Since DiracSmash is somewhat unusual [33,39] as a consequence of controlling and minimizing the number of independency assumptions that are inevitable features of inference net methods, note that this means the following. It is asked that the disease must be conditional upon the query represented by the so-called Hitlist of interdependent factors (interdependent on each other and the disease target) and the so-called Wishlist of factors independent of each other but interdependent with the above, plus further associated factors associated with all the above and found by data mining. This supervised by the Hitlist and Wishlist as query [33,39]. In this case, in most cases data was flagged as sufficient that all entries could be placed on the Hitlist. These Hitlist and Wishlist entries when necessary can be found by a preceding unsupervised data mining step [28,34]. The odds inference net performed slightly but significantly better than single hidden layer neural net. The DORs were respectively 52 vs. 35, 34 vs. 29, and 45 vs. 29, and the points at which accuracy = sensitivity = specificity that can be calculated from the DOR directly as well as assed at the training set [60] were 88% vs. 88.5%, 85% vs. 84%, and 87% vs. 84%. A study on renal failure predicted an LR of 295 with an accuracy = sensitivity = specificity point of 94.5%, but the neural net failed to converge over several days. Bleeding peptic ulcer predictions gave DOR of 64 versus 63 and 6.6 vs. 7.8 corresponding to 89% vs. 89% and 72% vs. 72%, so that in this case the glass and black methods performed comparably, with the neural net doing slightly better in one case. However, in a further run the glass box again failed to converge.

Studies were also carried out with Oxford Drug Design and Oxford University Department of Chemistry in drug discovery and development, of interest here not only because of the relevance to responding to emerging epidemics but also because of the opportunity it provided to compare sophisticated Machine Learning and Deep Learning DL models running with more resources [85]. The data comprised inhibitors of two

different enzymes with 203 chemical descriptors for each inhibitor, mainly compositional data such as the number of aromatic heterocycles in the compound. A third dataset for another enzyme performed well for DiracSmash in a blind test and thus opened interest in a joint study, but collaborator's ML and DL were not compared. The DiracSmash glass box approach gave (i) a DOR of 5.0 (69%) vs. compared with an average of 4.6 (62–72%) for ML and DL methods, and (ii) a DOR of 2.4 (61%) compared an average of 2.8 (60–65%) for ML and DL methods. In one study, DL performed comparably or slightly better than the glass box approach, though taking longer "training" time.

Several observations should be made on scalability. In the above chemical study the glass box method went through a stage (data mining) analogous to the training phase of black box methods much more rapidly than the black box methods, even though running in the background of normal office use on a T409s laptop without memory expansion, in contrast to mainframe resources. However, the *testing* phase of such glass box methods is in general slower. In a clinical setting of a patients, a run time of a few seconds is sufficient for a patient-physician clinical setting, but time consuming if it is to be run on hundreds of thousands of patients representing a test set, and a server cluster is required. This is not least because there may be many thousands or even several millions of probabilities involved in the full prediction process [33,39]. However, testing the explanatory simplified models that may be generated is extremely fast, and can be done on a personal laptop. For any one kind of computational platform and data, the time required scaling at least as 2^N for the number N of attributes, i.e. number of columns of structured data. For data considering much more than 200 attributes, feature selection by a first pass of unsupervised data mining, discussed frequently above, is indicated. However, all other factors being equal, the execution time is proportional to the number of records. The issue thus becomes simply a matter of managing Big Data in blocks, for example by writing code in SCALA to run on multiple processes in a SPARK environment as discussed briefly in Introduction Section 1.1. Where the number of medical records involved in the studies mentioned in this paper is only of the order of a million and typically 500,000 to 700,000, some 10 million older (ICD9) health insurance claims records (in many respects rather like health records) have been processed and led to the observations on scalability above. The present studies are in part preparation for access to some 100 million e.g. Truveta records, and this is not expected to be a problem with feature extraction which, recall, is inherently a glass box method. Based on the above, a key factor to resolve any prohibitive computation of the above type is feature selection. A typical biomedical data set of 200 attributes will run in time roughly proportional to $2^{200}{\sim}10^{60}$, and one of 100 attributes in time roughly proportional to $2^{100}{\sim}10^{30}$, that is, 10^{30} times faster. When needed, much of the sofware used here comprised algorithms that, at the cost of some assumptions, seek to limit such combinatorial explosions. Simple examples are elimination of combinations for which there is unlikely to be sufficient data *a priori*, and probabilistic sampling that is different for each record (including reiterative passes over the data) until convergence.

4.3. Studies on synthetic data for respiratory virus patients

In these studies pulmonary disease was considered a severe respiratory virus disease simply by changing the attribute names 'pulmonary disease' (with values yes/no) to 'serious respiratory virus disease' in preexisting clinical data. This has the advantage of involving real data with realistic distributions of multi-attribute (multifactor) probabilities. The reasonableness of this was verified by looking for symptoms as in Tables 3 and 4, and by comparing Q-UEL tags like those shown shortly below. For example, applying data mining to the above modified data showed partial pressure of oxygen low and carbon dioxide high, along with low blood glucose and low white blood cell count, with an Obwd (Likelihood Ratio) of 25.75. While the first three of these associated with respiratory stress from other causes in older data and with early COVID-19 wave 1 and wave 2 data, care must be taken with assumptions of cause and effect. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) proliferates in acidic blood that is possibly from other causes that magnifies the severity of COVID-19. Lower blood pH seems strong prognostic factor for fatal outcomes in critically ill COVID-19 patients. The low glucose appeared to more due to poor regulation of blood glucose in diabetes. It is understood that the virus can worsen complications of diabetes, as Tables 1 and 2 show, including hypoglycemia as well as severe hyperglycemia. In many patients the white blood cell count is raised as expected, but more severe COVID-19 is also found associated with patients with comorbidities and conditions such as autoimmune diseases such as rheumatoid arthritis, HIV and viral hepatitis, in other conditions of agranulocytosis and neutropenia when neutrophils are reduced, cancer treatment such as radiation, and even certain antipsychotic medications. Although quantitative agreement was neither needed nor expected, preliminary calculations for severe COVID-19 on the basis of comorbidities gave an expected LR of 20.5, in reasonable agreement with the above 25.75. They still remain, in practice, synthetic data. Also, low partial pressure of oxygen in arterial blood of less than 75 mm Hg and partial pressure of carbon dioxide greater than 45 mm Hg is rather artificially taken as analogous to criteria often used in COVID-19 studies (30 or more breaths/minute, a blood oxygen saturation less than 94% or less, and a ratio of the partial pressure of arterial oxygen to the fraction of inspired oxygen less than 300 mm Hg). The first step, where exact counting is possible, is analogous to the Truveta Studio, a primary interface for Truveta [54], and direct calculation from source data in general, as also done for Tables 1–4.In simulated data for severe symptoms of a future COVID-19 or avian influenza pandemic, results like the following were obtained.

Using chronic pulmonary disorder as a severe respiratory virus disease, the above normal partial pressure gave an LR of 9.88 and a PO of 0.78 using DiracSmash. However, that is based on a surprisingly high prevalence of the original condition of chronic pulmonary disease of 0.079 (7.9%) extensively from hospital record data that was likely by its nature biased. In the US overall, 6.4% of adults reported that they had been diagnosed with chronic obstructive pulmonary disease (COPD). However, in hindsight, more than 50% of adults with low pulmonary function were not aware that they had COPD so the actual number may be much higher [67]. Moreover, the US statistics vary 3.5–19.7% by county and 4.0–10.7% by state, with states along the Mississippi and Ohio rivers having the highest prevalence and some earlier estimates giving a ceiling of 13.6% for worst cases. While the current main but not sole cause is agreed to be smoking, these lung disorder figures do not bode well as comorbidities for a new respiratory virus pandemic.

In consequence, setting the prevalence to 5% severe respiratory disease that would give a PO of just 6.25 seems modest. The additional data mining given the above as required factors gave 1,786,404 Q-UEL tags used in prediction to give the above LR 13.4 and PO 1.04. An example of a tag for a patient not at risk from a drug abuse lifestyle and AIDS, where the PO corresponds to Ofwd:=0.2964 (uncorrected for 5%) and the LR to a moderate Obwd:=3.7483. Datamining was adjusted so that the factors following the 'with' operator could go up to five. However, note that data is already sparse: the overall prediction depends on combining the large number of tags.

< 'Serious respiratory virus disease':='Y' Pfwd:=0.2402 Ofwd:=0.2964 Efwd:=0.7814 | **if**:=(assoc:=3.2768, count:=4, factors:=(3,5)) | 'PO2 mmHg':='low' 'PCO2 mmHg':='high' **with** 'Drug abuse':='N' 'HIV and AIDs':='N' Pbwd:=0.0001036 Obwd:=3.7483 Ebwd:=9.8802 >

Two other arbitrary examples are as follows: the first shows a negative impact of abnormal electrolyte and blood sugar disorders, and the second shows abnormal pH, glucose, and white blood cell count test results. In both cases, time-stamp-related information such as first, second, and last test results have been removed for simplicity: like the above and Tables 1–4, these relate to conditions to first readings prior to infection and in the original data prior to diagnosis of chronic lung disease, i.e., they relate to susceptibility to severe disease.

< 'Serious respiratory virus disease':='eqY' Pfwd:=0.7163 Ofwd:=1.6593 Efwd:=0.7814 | if:=(assoc:=9.7732, count:=4, factors:=(3,6)) | 'PO2 mmHg':='low' 'PCO2 mmHg':='high' **with** 'PTAV PROTROMBIN TIME sec':='normal' 'K POTASSIUM mmol/L':='normal' 'normal' 'Na SODIUM mmol/L':='normal' Pbwd:=0.00010358 Obwd:=20.9802 Ebwd:=9.8802 >

< 'Serious respiratory virus disease':='eqY' Pfwd:=1.0000 Ofwd:=2.0368 Efwd:=0.7814 | **if**:=(assoc:=13.6443, count:=1, factors:=(3,6)) | 'PO2 mmHg':='low' 'PCO2 mmHg':='high' **with** 'PH:='low' 'GLUCOSE BLOOD mg/dL':='low' 'WBC WHITE BLOOD CELL k/uL':='low' Pbwd:=4.1688e-005 Obwd:=25.7542 Ebwd:=9.8802 >

Recalling the above LR of 9.88 and a PO of 0.78 for all the conditions being prior, it was surprising to find that *final* clinical laboratory test result readings taken gave LR of 10.59 and PO of 8.43, and 10.55 and 0.83 after data mining. This may represent a lack of therapeutic success in the original clinical data for chronic pulmonary disorder and, hence, indirectly, the effect of artificial data. Setting prior oxygen and carbon dioxide levels as both normal and the last as low and high, respectively, gave LR 10.12 and PO 0.80, respectively, and 11.35 and PO 0.90 after datamining 365,064 tags. Using much fuller descriptions of patients for severe respiratory virus disease with congestive heart failure, chronic kidney disease, and bleeding peptic ulcer did not change significantly for those even with chronic respiratory disorders, presumably here not so much because the data is artificial but because the conditions are already severe and close to the upper limits of prediction available for real-world data [39]. They were in the 70–90% range for congestive heart failure and, more recently, in the 80–90% range for renal failure [39]. Predictions for successful outcomes of therapies for bleeding peptic ulcer (see below) were typically 70%–90% (lower results reflected implausible choices of therapy for comparison). Because the methods are general, the following may be presented as added evidence as to the levels of prediction reachable. When applied to predict candidate drugs such as enzyme inhibitors, this is a particularly difficult but typical problem of having very few active compounds, accuracies lay in the range of 70–79%, sensitivities 50–70%, and specificity 70–80%, very similar to methods obtained by collaborators using Deep Learning, and superior to machine learning methods like Logistic Regression. For secondary structure prediction and fold motifs in proteins, it was possible to achieve accuracy sensitivities and specificities greater than 90% and up to close to 100% using training data comprising large collections of protein structures from which obvious homologous proteins were removed [38]. The above are studies analogous to using a training set and predicting on a test (validation) set. Details are being prepared for publication elsewhere for those studies whose results are not cited above. However, the immediate message is that using artificial data to try and emulate future pandemics is not a satisfactory pursuit, though at least it draws attention to issues on which to focus. Still, some data is needed from the emerging pandemic, even if it is rather early and sparse.

4.4. Comparison with other glass box methods

It would be implausible for authors of any review of many techniques to obtain and run all on the same data in order to make comparisons, but as explained below we do simulate, naturally with several caveats, some of the techniques available. This is possible because the flexibility of the main glass box methods used. The further problem with comparing the above studies on synthetic data is that most other approaches do not directly report many of the PO, LR, DOR, DOR*, and SS measures of interest here. As a basic first but important step, a comparison was made with a diverse set of structured unsupervised data mining [83], especially in the sense of unsupervised use here (Methods Section 3). QFANO [21], DiracMiner [28], and the unsupervised mode of DiracSmash [33, 39] are sufficiently flexible alone or in combination to provide or closely emulate a variety of kinds of structured mining for quantitative, qualitative, and mixed data. These included, but were not confined to, (a) mining dependence among components; (b) mining homogeneous or nonhomogeneous groups among samples; (c) mining dependences within local groups; and (d) mining topological structures with organized groups (akin to self-organizational maps); for definition of specific methods Ref [83] should be addressed. It cannot be claimed that the simulations used were algorithmically equivalent to any specific one of the above programs and fundamentally worked with mutual information and multi-factor association constants, and various conditional probabilities from which data properties analogously indirectly deduced, as well as noting negative associations and even combinations that should be seen as joint events but are not seen. QFANO can also perform multifactor Pearson correlation and express it as if it were an association analysis, meaning in the same format.

One method of particular interest is the Random Forest algorithm first defined by Ho [84] which, although often considered a black box machine learning method, has two features in common with the above essentially glass box approaches. First, it is concerned with making predictions as a basis for decisions or analysis, albeit it produces in one shot decision trees of several possible decision trees. Second, the trees generated are really a graphic representation of association clusters and subclusters and the random forest tree structure can be alternatively represented by association constants seen as "rules". DiracSmash can in effect be seen as a predicted optimal path down the decision tree and can be controlled to explore other branches. From data-mined association constants from DiracMiner [28] used to form the queries for DiracSmash that drive further data mining (partially supervised by the queries), the latter can generate simplified rule sets with optional levels of sophistication versus simplicity. These were of importance as explanatory rules guiding future synthesis of drug candidates described in Section 4.2, and perhaps even more closely resemble the rules by which random forest trees can be explained [84]. The advantage of formulating this in terms of a data mining approach like that above is typically greater efficiency, and particularly greater accuracy and reliability primarily due to fewer independency assumptions and greater control over them. However, this study is preliminary and it is arguable as to how closely it is to the Random Forest algorithm.

There are also some exceptions to the association-based analyses, notably pattern normality ranking and a PATFACTORS tool option within DiracSmash [33]. The latter relates to a variety of efficient pattern discovery methods by Rigoutsos and collaborators that have been developed from the original TEIRESIAS algorithm [86], and this can be emulated by a pattern discovery option in DiracSmash [33]. Even here, however, the approach is emulated rather than reproduced. The original TEIRESIAS algorithm and its variants are combinatorial in nature and able to produce all patterns that appear at least a (user-defined) minimum number of times and are very efficient by avoiding the enumeration of the entire pattern space; any reported pattern cannot be made more specific and still appear at the same positions within the input sequences. The analogous option in DiracMiner takes a column-number view and traditional arrays, as opposed to its usual procedure of working with attribute type names as keys to specify dictionary (associative, hash) arrays, and constructs a probability profile across patterns based on the number of times each attribute value shows up. However, the patterns are processed and reported on Q-UEL-PATFACTORS tags [33] with an appearance like Q-UEL tags shown in Section 3 Methods and later below. The approach is not intended to be as efficient as Rigoutsos' algorithms but to report and highlight the probabilistic nature of the pattern's components, but the final effect is similar.

Such studies are still under analysis and will be reported elsewhere, andare currently arguable as to the exact correspondence. However, it gives indication of the performance that class of algorithm and a general qualitative conclusion is that Deep Learning in particular behaved rather similarly to a glass box used here when used in predictive mode. The aspiration is not so much to supercede DL but to work with it to provide visibility, explainability, and probabilistic measures that are crucial for

present purposes. When extensive numbers of demographic and clinical factors were specified, or mining confined to records from that strata of the population, i.e., when there was plentiful information, all gave LRs of 50 or more and an accuracy = sensitivity = specificity point on a ROC curve [39,60,61] of 88% or more. QFANO, DiracMiner, and DiracSmash tended to be able to handle *high dimensional* data mining [87] more accurately routinely producing tags with 6 attributes as shown below, and in other studies on congestive heart failure and renal failure [33,39] 8 and even up to 11 coming together significantly more than would be expected on a chance basis.

Note that if a "joint event" of so many attributes is seen only 4 times in approximately a million records, then the expectation is that approximately 5 million records should be used to see it a more persuasive 20 times. An exception is pattern discovery by methods like TEIRESUIS and DiracSmash PATFACTORS, which can generate patterns with many attributes. Still, they face the same problem that the more attributes, the fewer the number of occurrences of the pattern. Arguably far more important than competing in terms of numbers of attributes is that the partially summated zeta function approach provides a Theory of Expected Information that allows many, perhaps millions of Q-UEL tags with very little supporting data (say 1 to 10 occurrences) to be brought together in prediction [33,39]. Recently it has been shown [60] that replacing the partially summated zeta function with an analogous standard logarithmic function in methods such as DiracSmash almost always does significantly worse in predictions than using the zeta function, and never better [60]. Evidently the situation is analogous to that in a court of law where a great deal of weak evidence can overthrow a decision made without it. To implement the above using large odds inference nets, DiracSmash, in more typical use, performs high dimensional data mining *supervised* by a first step involving a specified target, such as a disease and exact counting specified in a list of interdependent factors called the Hitlist, plus a list of factors called the Wishlist that are interdependent with target and Hitlist but assumed mutually independent [33,39]. Ingine's DiracMiner [28] can provide the associations that can automatically populate the above lists using association constants showing degrees of interdependency but for DiracBuilder, the adjoint conditional probabilities of general form $P(A|B)$ and $P(B|A)$. While variants of DiracSmash, such as QuickPredict [39], can use tags both from DiracMiner and previous DiracSmash runs, the focus of DiracSmash, as described above, is in terms of Predictive Odds and Likelihood Ratios (LR). The LR is important in the single-patient clinical setting because if exceeding one, they predict the target, say a diagnosis or best therapy choice.

4.5. Considerations of DOR*

The DOR* is a recent but important development, not least because it appears to be a sensitive measure for predicting those therapies that produce good as opposed to poor outcomes. While the DOR* discussed in Theory and Methods, Section 2 continues to be important for present purposes and perhaps represents the most important measure because it is applicable to vaccine effectiveness and antiviral drug therapy, it is not included in the Tables because its role belongs in an even less predictable future than matters considered in Section 4.2 above. The Diagnostic Odds Ratio DOR that would be obtained by a predictive method used analytically corresponds closely to the LR in these studies because of the number and nature of the conditions. From the LR obtained conditional on a good outcome divided by that for a poor outcome, one likely obtains an even more reliable measure of the DOR*. However, the prevalence of the two kinds of outcomes depends on the therapies applied, the vaccines and therapeutic drugs, and the diagnostics and methods of primary prevention (See Discussion and Conclusions Section 5.2).

It is possible to have some sense of the values obtainable for DOR* for a future serious pandemic by the device shown above, which replaces chronic pulmonary disease with severe respiratory virus disease. In that case, a good outcome for any treatment applied is that partial pressures for oxygen and carbon dioxide are initially low and high, respectively, and finally normal. That gives a DOR of 8.35 for a good outcome. Changing the last two measures to stay as low oxygen and high carbon dioxide partial pressures gave a DOR of 9.59, the ratio giving a DOR* of 0.87. There was ample data for prediction, with 1,532,894 being generated and used in inference net construction in the first case and 9,326,934 in the second. This suggests that somewhat less than 50% of cases recover, which fits into the view that it is not currently possible to cure or reverse COPD completely. However, treatments and lifestyle changes can reduce its impact. Indeed, the word *chronic* in "chronic pulmonary disease" does, of course, suggest values should stay abnormal. The ability to cure pulmonary disorders is varied, and hospital admissions often relate to "bursts" of the preexisting condition that can be treated. The available data borrowed from the first wave of COVID-19 as pertinent to a new emergent pandemic did not include therapies also because they were limited for the first wave, the most obvious and successful for very severe cases being extracorporeal membrane oxygenation (ECMO) for severe respiratory distress, i.e., a ventilator. By supporting the heart and lungs, the ECMO machine stabilizes patients, giving their bodies more time to fight the virus. Here, the SS is an important measure for assessing the demand for ventilators. The absence of further treatment details regarding therapies and demographic and clinical data limited analysis, but some 84 % of patients over 70 years old died in the hospital, while 67% of patients aged 70 or less recovered, suggesting DOR*s of roughly 3–4 for younger patients and 0.2 for over the 70s given, of course, the condition that the patients were seriously ill. A DOR* of around 1 would indeed mean that the treatment has an approximately 50:50 chance of being effective. Overall, some 57% of COVID patients on ventilators survived hospital discharge, while estimates from the above synthetic study suggest 47%.

Although data is absent for the above viruses, DOR* has been applied to diagnosis, prediction of future occurrence, and best therapy for several diseases using real data [61]. These primarily involve DORs for congestive heart failure, chronic kidney disease, renal failure, and bleeding peptic ulcer. Generally, 8–14 is found for known effective therapies for diseases that are not very serious morbidities and do not have a very high fatality rate, and 4–5 for novel therapies for therapies that are plausible for significantly different demographic and clinical conditional factors studied. A DOR* of 8–10 is typical for known successful therapies or combinations of therapies, and that may be underestimated if one considers recurrence of the disease a separate case. For example, for bleeding peptic ulcer, a complicated case not least because of recurrence, a DOR* of 8.32 was obtained for the use of therapeutic upper gastrointestinal endoscopy with a proton pump inhibitor and without an H2 inhibitor (both reducing gastric acidity but with the former more potent) and when aldehyde dehydrogenase levels were initially high and normal treatments. Recalling that a DOR* of around 1 would mean that the treatment has an approximately 50:50 chance of being effective, a low score in the case of respiration for COVID-19 may still be useful in some circumstances, especially when combined with other therapies. An analogy might be the patient with a bleeding peptic ulcer who may be using a proton pump inhibitor before, during, and after therapeutic endoscopy, so the drug used in less serious cases may well prove effective before a scheduled therapeutic endoscopy appointment. This unpleasant procedure carries some risks of perforation and is not always effective.

5. Discussion

5.1. Limitations

There has recently been an explosion in AI methods, primarily black box, that at first glance would appear to pose no difficulties providing that here is resource to address at least the order of 100 million medical records. For methods that are not combinatorically explosive or a notion of convergence is not appropriate, coving all US medical records seems

plausible [25]. While the focus in this review has been primarily on glass box AI of the type that involves data mining at least at some stage, these and black box approaches have been discussed, and all suffer from one main difficulty. Even in large numbers or regularly updated medical records, relevant data due to the action of the new pathogen is likely to be extremely limited in the first few days of a respiratory virus epidemic. The nature of the studies carried out in this review therefore become even more relevant: one is setting expectations based on previous epidemics and general performance of predictive techniques in a medical setting. Such a study also, as shown in the tables, identifies features to which attention should specifically directed. Indeed, this can form the basis of feature selection, inevitably a method of at least some glass box character, to limit the high dimensionality of the data. There is no doubt, however, that techniques appropriate to managing sparse data are important [28] and that is why much attention has been given to the partially summated zeta function here. It is often considered that valid results in data mining require adequate "support", i.e. be seen several times, but the real notion of support should be the association constant K, indicating the extent that the number of observations differs from what is expected on a random basis. Indeed, "negative associations" ($K < 1$) are important in medicine (e.g. we wish negative association between prevention and disease) and an extreme case is when attributes should be seen together on the basis of individual probabilities (i.e. are "existentially qualified") but are not (e.g. pregnant patients born male). This is not too difficult to estimate for pairs, but the problem is combinatorically explosive or pairs of attributes, and more algorithmic research would be useful. Moreover, it is clear that a more formal approach should be possible in which the earlier studies represent prior knowledge, and where they are glass box methods that work with probabilities and odds, application of Bayes rule, based on prior probability times likelihood, should be fairly straightforward. The use of virtual frequencies [28,33,39] outlined briefly in Results Section 4.1, is an equivalent notion that combines with the zeta approach to sparse data. However, further study on this would be useful.

As regards comparison of methods to see where improvements might be made, glass box algorithms tend to be much more efficient in the step analogous to training, with only a few seconds or minutes per million records given reasonable server resources, but when the method is elaborate with many probabilities, less efficient at testing on large numbers of records. Certainly, speed of use once trained is a feature of many black box algorithms, but significant computing power may be required. However, an important step in responding to an epidemic would be to apply the techniques to the relatively few records or data for patients with the disease that quickly become more abundant after the earlier stages. As to comparisons between glass box methods, the key differentiator determining accuracy is the number of assumptions made, which typically means the number of independencies assumed between attributes and whether those attributes are chosen judiciously without unnecessary presumptions of causality. This applies obviously for black box methods, but it is difficult to assess what interdependencies and independencies are implied in the box. Explainable AI needs also to focus on that aspect.

5.2. The DOR* measure

The DOR* measure appears to be novel although, because the concept is simple, it may have been addressed before, at least by looking at two values, one for good and one for poor outcome. The DOR* is not included in Tables 1–4 because its role belongs in the less predictable future. The rather invariant results for different outcomes for a synthetic model may suggest that either the model is not a good one (say, particularly with chronic pulmonary disease reinterpreted as severe respiratory virus disease) or that the DOR* measure is inappropriate. The latter is unlikely in view of the relatively simplicity and natural character of the measure, which appears formally correct when expressed in terms of information measures as in the equations of this paper, but it is of course important that other attributes mentioned are appropriate, the diseases well defined, and the basis for estimating good and poor outcome is adequate. An automatic method like DiracMiner takes care of appropriateness of the other attributes by considering all and assessing their contribution, but the other two issues involve "definitions of state", currently matters for the user. However, the true success rate from ventilation of COVID patients was not very far from the predictions of good and poor outcomes, and tests on other diseases in entirely real data gave reasonable results. As also noted, the prevalence of the two kinds of outcomes depends on the therapies applied, the vaccines and therapeutic drugs available, and the diagnostics and methods of primary prevention. It will, of course, also depend on the pathogen involved, and with an improved understanding of it, recall that the reported case fatality rate in the first wave of COVID-19 declined by about 83.3% in the second wave.

Nonetheless, it would be better to say that DOR* is not explicit in the tables rather than missing. Essentially, one wishes to choose preventative, curative, and ameliorative methods (commonly called primary, secondary, and tertiary "preventions" by epidemiologists) such that all SS measures are reduced as much as possible. The measures presented here may provide a baseline for that endeavor, at least qualitatively. We may also expect that comorbidities and unfair equity issues such as deprivation and perhaps ethnicity might, unfortunately, still lead to higher measures, in which case one should also strive to reduce the SS values involved.

The above implies the useful role of the Likelihood Ratio LR, DOR*, and SS measures in driving and assessing clinical trials for vaccines and therapeutics against the emerging pandemic. The association constant is also useful. All can be calculated from the output of glass-box methods [33,39], and all are specified on Q-UEL tags for specific "clusters" of demographic and clinical factors, as illustrated above. An important aspect of the analysis of many freshly updated patient records is that they enable the production of meaningful measures and predictions that can be made not just for the future but for current analytical purposes. But when used to make predictions for the near future, this includes helping design clinical trials and certain types of prospective cohort studies [58]. The need for a trial in a new pandemic arises early. A Phase 1 clinical trial evaluating an investigational vaccine designed to protect against COVID-19 began at Kaiser Permanente Washington Health Research Institute in Seattle. It was scheduled to enroll 45 healthy adult volunteers ages 18–55 years over approximately 6 weeks. The first participant received the investigational vaccine on March 16, 2020. These are, as is not uncommon for clinical trials, small numbers, unrepresentative of the large population to which results will, ultimately, be applied, and assigning attribution of the effects can be difficult without prior studies [61].

5.3. Further considerations

From the experimental study viewpoint, a purely computational retrospective study using many patient records has the additional advantage that one component of the trial can be halted. In contrast, other components continue, essentially viewing the factorial design as separate trials for each factor. Ingine's glass-box methods facilitate this [33,39]. For a new experimental trial, when considering the termination of one component of the trial, an evaluation of the effect on statistical power is critical. The termination of one component will reduce the power to detect interactions. It will complicate analyses and subsequent interpretations of the main effects and interactions [61]. When likely interactions, such as positive or negative associations, are already understood from a retrospective cohort study using many records, the researcher and manager of an experimental trial are in a better position to make decisions.

Participant recruitment is said to be more complex in factorial trials and can decrease accrual rates [61]. Study participants must not only be willing but meet criteria for treatment with each intervention with no

contraindications to any of the possible treatment combinations. and have a willingness to consent to all interventions and procedures. Association analysis in a retrospective cohort study of many records could facilitate procedure because adherence to protocol can also be complicated due to the multiple interventions and interaction effects should always be evaluated even if the trial design assumed no interaction. Association analysis and predictive models can evidently provide helpful measures and a transparent summary of each treatment cell. "Transparency" resonates with a "glass-box" rather than "black-box" machine learning and deep learning methods. Researchers need to be aware of the multiplicity issue given an assessment of multiple interventions and impacting factors. Still, researchers typically control the error rate for the assessment of each factor separately rather than controlling a trial-wide error rate; when interactions exist, then it is inappropriate to interpret single global intervention effects, and one must estimate intervention effects conditional upon whether the other intervention is provided using subgroup analyses [61]. There could be two effects of intervention A: one for patients that receive intervention B and one for patients that do not receive intervention B. This is an example of multifactor association and probability analysis, a relatively simple one in this case, e.g., P(outcome | A, B) and P(outcome | A, not B), or corresponding Likelihood Ratio P(outcome, A |B)/P(outcome, A | not B). Recall that a ratio of the latter LR for a good outcome versus that for a poor outcome is usually a good estimate of DOR*.

6. Conclusions

Because a review is based on history and a position paper is based on supposition, we cannot be certain how effective the principles considered in this review and position paper will be in the case of a very new dangerous strain of COVID-19 or a human Avian Influenza pandemic. In particular, we might ask how the therapies, vaccines, and therapeutic drugs available affect the prevalence of different outcomes, and whether we can seek a quantifiable way to measure their impact within the models. These important considerations deserve extensive discussion beyond current scope, but designs of diagnostics, vaccines, and peptidomimetic lead drugs based on comparing early SARS virus and many other coronaviruses genomes, along with considering repurposing of drugs of interest that were of interest in the response to the earlier SARS outbreak (e.g. emodin) were major points of interest in Ingine's COVID-19 papers [10,12–16]. It is plausible that similar extension from past COVID to future more troublesome strains, and from past porcine and human influenzas and bird-to-human cases to human to human avian flu transmission, will be important. Indeed in above the COVID-19 studies [10,12–15], comparison between related viruses and their impact on humans was, as is so often the case, of great importance. Information concerning the molecular character for an unfamiliar pathogen and its first impact is hugely reduced in value if similar causative agents and effects are not found elsewhere in human history or animals, and the similarities and differences used to form a response. The new problem cannot stand alone. The alternative would be an understanding of humans and the human body's response to a new pathogen on a theoretical, *ab initio*, molecular scale basis that is still beyond our current capabilities.

Still, we have learned quite a lot. After a period of complacency, AIDS [1], mad cow disease, threats of Ebola and porcine influenza outbreaks, and then SARS and COVID-19 have shown both mistakes to avoid and processes to follow. The lessons from history are not perfect: as noted in the financial times of December 6th, 2023, the UK government was slow to respond to COVID-19 because predictions from worst-case scenario modeling for mad cow disease and swine flu had not been realized, while COVID-19 the consequences turned to be disastrous [68]. They may be more disastrous for a radically new COVID strain. They are almost certainly to be far more disastrous for avian influenza and perhaps for something entirely unexpected. Consequently, we have hopefully learned to assume the worst-case scenario until there is hard evidence to the contrary. But with that cautiously in mind and recognizing that we are more alert to the risks of COVID and avian influenza, access to large numbers of digital patient records, ideally regularly updated, will be enormously helpful.

CRediT authorship contribution statement

B. Robson: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **O.K. Baek:** Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

There is not believed to be any significant https://www.sciencedirect.com/science/article/abs/pii/S0010482524001690. Professor Barry Robson reports a relationship with Ingine Inc. that includes: equity or stocks. Dr. O. K. Baek was in receipt of a grant of the South Korean government, as cited in the paper, for which purpose BR is a contractor through Ingine Inc.

Acknowledgements

This project received support from a South Korean government grant, managed by Dr OK Baek, specifically the Electronics and Telecommunications Research Institute (ETRI) grant 23ZS1100, focused on advancing Core Technology Research for Self-Improving Integrated Artificial Intelligence Systems. Author Barry Robson acknowledges the support of Ingine Inc. and collaborators in the cited papers who have contributed to some of the ideas expressed here that helped review the background.

References

[1] Garrett L. The coming plague. In: Newly emerging diseases in a world out of balance. Farrar, Straus and Giroux; 1994.
[2] Tyrrell DA, Bynoe ML. Cultivation of viruses from a high proportion of patients with colds. Lancet 1966;1:76–7. https://doi.org/10.1016/s0140-6736(66)92364-6. . [Accessed 1 November 2023].
[3] Hamre D, Procknow JJ. A new virus isolated from the human respiratory tract. Proc Soc Exp Biol Med 1966;121:190–3. https://doi.org/10.3181/00379727-121-30734. . [Accessed 1 November 2023].
[4] Masters PS. The molecular biology of coronaviruses. Adv Virus Res 2006;66:193–292. https://doi.org/10.1016/S0065-3527(06)66005-3. . [Accessed 1 November 2023].
[5] https://www.bbc.co.uk/news/world-55756452 (last accessed January 1 2022).
[6] Walker PGT, et al. The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. Science 2020;369:413–22. https://doi.org/10.1126/science.abc0035.
[7] Robson B. Preliminary bioinformatics studies on the design of synthetic vaccines and preventative peptidomimetic antagonists against the Wuhan Seafood Market coronavirus. Possible importance of the KRSFIEDLLFNKV motif. 2020. https://doi.org/10.13140/RG.2.2.18275.09761. Epub 30th January on ResearchGate, . [Accessed 1 November 2023].
[8] Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579(7798):270–3. https://doi.org/10.1038/s41586-020-2012-7. Epub 2020 Feb 3, . [Accessed 1 November 2023].
[9] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020; Feb 22;395(10224):565–74. https://doi.org/10.1016/S0140-6736(20)30251-8. Epub 2020 Jan 30, . [Accessed 1 November 2023].
[10] Robson B. Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus. Comput Biol Med 2020. https://doi.org/10.1016/j.compbiomed.2020.103670. Epub 2020 Feb 26: 103670, . [Accessed 1 November 2023].
[11] Xia S, Liu M, Wang C, Xu W, Lan Q, Feng S, Qi F, Bao L, Du L, Liu S, Qin C, Sun F, Shi Z, Zhu Y, Jiang S, Lu L. Inhibition of SARS-CoV-2 (previously 2019-nCoV)

infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. Cell Res 2020;30 (4):343–55. https://doi.org/10.1038/s41422-020-0305-x. Epub 2020 Mar 30, . [Accessed 1 November 2023].
[12] COVID-19 coronavirus spike protein analysis for synthetic vaccines, a peptidomimetic antagonist, and therapeutic drugs, and analysis of a proposed Achilles' heel conserved region to minimize probability of escape mutations and drug resistance. Comput Biol Med 2020;121:103749. Epub 2020 Apr 11.
[13] Robson B. Bioinformatics studies on a function of the SARS-CoV-2 spike glycoprotein as the binding of host sialic acid glycans. Comput Biol Med 2020;122: 103849. https://doi.org/10.1016/j.compbiomed.2020.103849. Epub Jun 8, (2020), . [Accessed 1 November 2023].
[14] Robson B. The use of knowledge management tools in viroinformatics. Example study of a highly conserved sequence motif in Nsp3 of SARS-CoV-2 as a therapeutic target. Comput Biol Med 2020;202:125. https://doi.org/10.1016/j.compbiomed.2020.103963. Epub 2020 Aug, (2020), . [Accessed 1 November 2023].
[15] Robson B. Techniques assisting peptide vaccine and peptidomimetic design. Sidechain exposure in the SARS-CoV-2 spike glycoprotein. Comput Biol Med 2020; 128(2021):104124. https://doi.org/10.1016/j.compbiomed.2020.104124. Epub Nov 21, . [Accessed 1 November 2023].
[16] Robson B, Fishleigh RV, Morrison CA. Prediction of HIV vaccine. Nature 1987;4 (325):395. https://doi.org/10.1038/325395a0. . [Accessed 1 November 2023].
[17] Fishleigh RV, Robson B, Mee R. P. Fragments of prion proteins, US Patent 5,773,572.
[18] Robson B. From Zika to flu and back again. In: CAVIRC (Caribbean anti-virus informatics research center); 2016. https://doi.org/10.13140/RG.2.1.5000.6808. . [Accessed 1 November 2023].
[19] Robson B. Towards faster response against emerging epidemics and prediction of variants of concern. Inform Med Unlocked 2022;31:100966.
[20] Executive office of the president president's council of advisors on science and technology, report to the president realizing the full potential of health information technology to improve healthcare for Americans: the path forward. 2010. GOVPUB-PREX23-PURL-gpo2425.pdf (govinfo.gov), . [Accessed 15 January 2024].
[21] Robson B, Caruso TP, Balis UGJ. Suggestions for a web based universal exchange and inference language for medicine. 1 Comput Biol Med 2013;43(12):2297–310. https://doi.org/10.1016/j.compbiomed.2013.09.010. Epub Sep 20, (2013), . [Accessed 1 November 2023].
[22] Robson B. Towards new tools for pharmacoepidemiology. Adv Pharmacoepidemiol Drug Saf 2013;1:6. https://doi.org/10.4172/2167-1052.100012. . [Accessed 1 November 2023].
[23] Robson B. The new physician as unwitting quantum mechanic: is adapting Dirac's inference system best practice for personalized medicine, genomics, and proteomics?. 8 J Proteome Res 2017;7(6):3114–26. https://doi.org/10.1021/pr070098h. Epub 2007 Jul 3, . [Accessed 1 November 2023].
[24] Robson B, Caruso TP. A universal exchange language for healthcare. In: Lehmann CU, Ammenwerth E, Nohr C, editors. Stud health technol inform, vol. 192. IOS Press; 2013. p. 949.
[25] Robson B, Caruso TP, Balis UGJ. Suggestions for a web based universal exchange and inference language for medicine. Continuity of patient care with PCAST disaggregation. Comput Biol Med 2014;56:51–66.
[26] Robson B. Hyperbolic Dirac Nets for medical decision support. Theory, methods, and comparison with Bayes Nets 2015;51(2018):183–97. https://doi.org/10.1016/j.compbiomed.2014.03.014. . [Accessed 1 November 2023].
[27] Deckelman S, Robson B. Split-complex numbers and Dirac bra-kets. Commun Inf Syst 2015;14(3):135–49. https://www.academia.edu/33231629/Split-complex_numbers_and_Dirac_i_bra-kets_i_. [Accessed 1 November 2023].
[28] Robson B, Boray S. Implementation of a web based universal exchange and inference language for medicine. Sparse data, probabilities, and inference in data mining of clinical data repositories. Comput Biol Med 2015;66(2015):82–102. https://doi.org/10.1016/j.compbiomed.2014.10.022. Epub Nov 4.
[29] Robson B, Boray S. Interesting things for computer systems to do: keeping and data mining millions of patient records, guiding patients and physicians, and passing medical licensing exams, Bioinformatics and Biomedicine (BIBM). In: Proceedings 2015 IEEE international conference, vol. 21015l; 2015. p. 1397–404.
[30] Robson B. Boray, data-mining to build a knowledge representation store for clinical decision support. Studies on curation and validation based on machine performance in multiple choice medical licensing examinations. Comput Biol Med 2016 Feb 26;73(2015):71–93. https://doi.org/10.1016/j.compbiomed.2016.02.010. 2016, . [Accessed 1 November 2023].
[31] Robson B. Studies in using a universal exchange and inference language for evidence based medicine. Semi-Automated Learning and Reasoning for PICO Methodology, Systematic Review, and Environmental Epidemiology 2016;79: 299–323. https://doi.org/10.1016/j.compbiomed.2016.10.009. Epub Oct 17, (2016), . [Accessed 1 November 2023].
[32] Robson B, Boray S. Studies of the role of a smart web for precision medicine supported by biobanking. Per. Med. 2016;13(4):361–80. https://doi.org/10.2217/pme-2015-0012. Epub Jul 5, (2016).
[33] Robson B, Boray S. Studies in the extensively automatic construction of large odds-based inference networks from structured data. Examples from medical, bioinformatics, and health insurance claims data. Comput Biol Med 2018;95: 147–66. https://doi.org/10.1016/j.compbiomed.2018.02.013. Epub Mar 21, (2018), . [Accessed 1 November 2023].
[34] Robson B. Bidirectional General Graphs for inference. Principles and implications for medicine. Comput Biol Med 2019;10(2019):382–99. https://doi.org/10.1016/j.compbiomed.2019.04.005. Epub 2019 Apr 13, . [Accessed 1 November 2023].
[35] Robson B. Extension of the Quantum Universal Exchange Language to precision medicine and drug lead discovery. Preliminary example studies using the mitochondrial genome. Computers in Biology and Medicine, Feb 2020;117: 103621. https://doi.org/10.1016/j.compbiomed.2020.103621. Epub Jan 20 (2020), . [Accessed 1 November 2023].
[36] Robson B. Computers and preventative diagnosis. A survey with bioinformatics examples of mitochondrial small open reading frame peptides as portents of a new generation of powerful biomarkers. Comput Biol Med 2019;140:105116. https://doi.org/10.1016/j.compbiomed.2021.105116. . [Accessed 1 November 2023].
[37] Robson B, Boray S. Studies in the use of data mining, prediction algorithms, and a universal exchange and inference language in the analysis of socioeconomic health data. Comput Biol Med 2019;112(2019):103369. https://doi.org/10.1016/j.compbiomed.2019.103369. Epub 2019 Jul 25, . [Accessed 1 November 2023].
[38] Robson B. Testing machine learning techniques for general application by using protein secondary structure prediction. A brief survey with studies of pitfalls and benefits using a simple progressive learning approach. Comput Biol Med 2012; 2021(138):104883. https://doi.org/10.1016/j.compbiomed.2021.104883. Epub 2021 Sep. 23, . [Accessed 1 November 2023].
[39] Robson B, Boray S, Weisman J. Mining real-world high dimensional structured data in medicine and its use in decision support. Some different perspectives on unknowns, interdependency, and distinguishability. Comput Biol Med 2022;141: 105118. https://doi.org/10.1016/j.compbiomed.2021.105118. . [Accessed 1 November 2023].
[40] Ahsan H, Sonn JK, Lee YS, Islam SU, Khalil SK. Keeping SARS-CoV-2 reinfections at bay, genetic engineering and biotechnology news. December 14. 2021. https://www.genengnews.com/immunology/keeping-sars-cov-2-reinfections-at-bay/. [Accessed 1 November 2023].
[41] Ahsan H, Sonn JK, Lee YS, Islam SU, Khalil SK. An overview about the role of adaptive immunity in keeping SARS-CoV-2 reinfections at bay. Viral Immunol 2021;34(9):588–96. https://doi.org/10.1089/vim.2021.0017. Epub 2021 Jun 8, . [Accessed 1 November 2023].
[42] SARS-CoV-2 and ACE2 receptor: combination of molecular dynamics simulation and density functional calculation. J Chem Inf Model 2021;61(9):4425–41. Epub 2021 Aug 24.
[43] Chen C, Boorla VS, Banerjee D, Chowdhury R, Cavener VS, Nissly RH, Gontu A, Boyle NR, Vandegrift K, Nair MS, Kuchipudi SV, Maranas CD. Computational prediction of the effect of amino acid changes on the binding affinity between SARS-CoV-2 spike RBD and human ACE2. Proc Natl Acad Sci U S A 2021;118(42): e2106480118. https://doi.org/10.1073/pnas.2106480118. . [Accessed 1 November 2023].
[44] Gürsoy E, Kaya Y. An overview of Deep Learning techniques for COVID-19 detection: methods, challenges, and future works. Multimed Syst 2023;29(3): 1603–27. https://doi.org/10.1007/s00530-023-01083-0. Epub Mar 25, (2023).
[45] Pinkstone, J., First UK case of swine flu variant found in humans. New strain of a virus contracted by 50 people globally has been found in North Yorkshire patient, https://www.telegraph.co.uk/news/2023/11/27/first-uk-case-of-new-pig-flu-found-in-humans-ukhsa-confirms/(last accessed November 1, 2023).
[46] https://netec.org/2023/01/30/situation-report-highly-pathogenic-avian-influenza-ah5n1/(last accessed November 1, 2023).
[47] https://www.everydayhealth.com/bird-flu/guide/(last accessed November 1, 2023).
[48] https://www.who.int/docs/default-source/wpro—documents/emergency/surveillance/avian- influenza/ai_20230224.pdf. (last accessed November 1, 2023).
[49] https://www.gavi.org/vaccineswork/five-things-know-about-whether-h5n1-bird-flu-outbreak-could-turn-pandemic (last accessed November 1, 2023).
[50] Yong E. Scientists create hybrid flu that can go airborne. Nature 2013;10(1038): 1476–4687. https://doi.org/10.1038/nature.2013.12925.
[51] https://www.newscientist.com/article/mg19826501-400-will-a-pandemic-bring-down-civilisation/(last accessed November 1, 2023).
[52] https://www.bbc.com/future/article/20181120-what-if-a-deadly-influenza-pandemic-broke-out-today (last accessed November 1, 2023).
[53] https://en.wikipedia.org/wiki/The_Postman_(film) (last accessed November 1, 2023).
[54] https://www.truveta.com/(last accessed November 1, 2023).
[55] Smits PD, Gratzl S, Simonov, Nachimuthu SK, Goodwin BM, Cartwright D, Wang MD, Baker C, Rodriguez P, Bogiages M, Althouse BM, Stucky NL. Risk of COVID-19 breakthrough infection and hospitalization in individuals with comorbidities. Vaccine 2023;41:15.
[56] Robson B, Baek OK. An ontology for very large numbers of longitudinal health records to facilitate data mining and machine learning. Inform Med Unlocked 2023;38:101204.
[57] Snigdha S, Aithal YR. Evaluation of deep learning models for identification and prediction of new strains of COVID-19. In: Bhattacharyya S, et al., editors. Computer intelligence against pandemics: tools and methods to face new strains of COVID-19. Berlin, Boston: De Gruyter; 2023. p. 257–86. https://doi.org/10.1515/9783110767681-013.
[58] Evans SR. Fundamentals of clinical trial design. Journal of Experimental Stroke and Translational Medicine 2010;3(1):19–27.
[59] Shah, P., Kendall, F., Khozin, S., Goosen, R., Hu, J., Laramie, J., Ringel, M., and Schork, N, Artificial intelligence and machine learning in clinical development: a translational perspective, Nature P. J. Digital Medicine, 69, 2(1), 2398-6352, (2029). https://doi.org/10.1038/s41746-019-0148-3.

[60] Robson B, Baek OK. Use of a theory of expected information for sparse data and adverse events in clinical trials and other biomedical studies. 2023. submitted for publication(still in review).
[61] Robson B, Baek OK. Glass box machine learning for retrospective cohort studies using many patient records. The Complex Example of Bleeding Peptide. Ulcer 2023. https://www.sciencedirect.com/science/article/abs/pii/S0010482524001690.
[62] Evans SR. Clinical trial structures. Journal of Experimental Stroke and Translational Medicine 2010;3(1):8–18.
[63] Priebe HJ. Problems of subgroup analysis in randomized controlled trial. BMC Anesthesiol 2020;20:186.
[64] Hemple S. The strange case of the broad street pump. University of California Press; 2020.
[65] Robson B, Baek OK. The engines of Hippocrates. From the dawn of medicine to medical and pharmaceutical informatics. Wiley; 2009.
[66] Barron E, Bakhai C, Kar P, Weaver A, Bradley D, Ismail H, Knighton P, Holman N, Khunti K, Satta r N, Wareham NJ, Young B, Valabhji J. Associations of type 1 and type 2 diabetes with COVID-19-related mortality in England: a whole-population study. Lancet Diabetes Endocrinol 2020;8:813–22. https://doi.org/10.1016/S2213-8587(20)30272-2. Epub 2020 Aug 13.
[67] https://www.cdc.gov/copd/basics-about.html (last accessed 14th December 2023).
[68] Hughs L, Gross A. Boris Johnson admits UK 'vastly underestimated' COVID threat. Financial Times; 2023. 6th December, https://www.ft.com/content/f93bd1ed-2936-40e5-b222-550232b695a6. [Accessed 14 December 2023].
[69] Speaker SL, Moffatt C. The National Library of Medicine Global Health Events web archive, coronavirus disease (COVID-19) pandemic collecting. J Med Libr Assoc : JMLA 2020;108(4):656–62. https://doi.org/10.5195/jmla.2020.1090.
[70] https://doi.org/10.1136/bmjgh-2020-004100.
[71] Tang K, Gaoshan J, Ahonsi B, et al. Sexual and reproductive health (SRH): a key issue in the emergency response to the coronavirus disease (COVID- 19) outbreak. Reprod Health 2020;17:59. https://doi.org/10.1186/s12978-020-0900-9.
[72] Walkey AJ, Kumar VK, Harhay MO, Bolesta S, Bansal V, Gajic O, Kashyap R. The viral infection and respiratory illness universal study (VIRUS): an international registry of coronavirus 2019-related critical illness. Critical care explorations 2020;2(4):e0113. https://doi.org/10.1097/CCE.0000000000000113.
[73] Nguyen TV, Tran QD, Phan LT, Vu LN, Truong DTT, Truong HC, Le TN, Vien LDK, Nguyen TV, Luong QC, Pham QD. In the interest of public safety: rapid response to the COVID-19 epidemic in Vietnam. BMJ Glob Health 2021;6(1):e004100.
[74] Ghozali MT. Implementation of the IoT-based technology on patient medication adherence: a comprehensive bibliometric and systematic review. Journal of Information and Communication Technology 2023;22(4):503–44. https://doi.org/10.32890/jict2023.22.4.1.
[75] Straus SE, Glasziou P, Tichardson WS, Haynes RB. Evidence-based medicine. How to practice and teach EBM. Elsevier, 5th Edittion; 2018.
[76] Levy A, Sobolev B, editors. Comparative effectiveness research in health Services. Springer; 2018.
[77] Glasa AS, Lijmerb GG, Prinsc MH, Bonsled GJ, Bossuyta PMM. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 2003;56:1129–35.
[78] Rohekar RY, Nisimov S, Gurwicz Y, Koren G, Novik G. Constructing deep neural networks by bayesian network structure learning. In: 32nd conference on neural information processing systems. Montréal, Canada: NeurIPS 2018; 2018. https://proceedings.neurips.cc/paper_files/paper/2018/file/95d309f0b035d97f69902e7972c2b2e6-Paper.pdf.
[79] Fort S, Hu H, Lakshminarayanan B. Deep ensembles: a loss landscape perspective. 2020. https://doi.org/10.48550/arXiv.1912.02757. arXiv:1912.02757vol. 2.
[80] Benois-Pineau J, Bourqui R, Petkovic D, Quénot G, editors. Explainable deep learning AI. Academic Press; 2023.
[81] Barraza JF, Droguett EL, Martins MR. Towards interpretable deep learning: a feature selection framework for prognostics and health management using deep neural networks. Sensors 2021;21(17):5888.
[82] Landolsi MY, Haoua L, Romdhane LB. Information extraction from electronic medical documents: state of the art and future research directions. Knowl Inf Syst 2023;65:463–516.
[83] Xu L. An overview on unsupervised learning from data mining perspective. In: Advances in self-organising maps. London: Springer; 2001.
[84] Ho TK. Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, vol. 1; 1995. p. 278–82.
[85] Robson B., Cooper C. The Importance of Using Complementary Glass Box and Black Box Machine Learning Approaches to Exploit Compositional Descriptors of Molecules in Drug Discovery and Aid the Medicinal Chemist. 2023. submitted for publication. In review.
[86] I Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. Bioinformatics 1998;14(1):55–67.
[87] Wang W, Yang J. Mining high-dimensional data. In: Maimon O, Rokach L, editors. Data mining and knowledge discovery handbook. Boston, MA: Springer; 2005.