Computers in Biology and Medicine 112 (2019) 103369

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Studies in the use of data mining, prediction algorithms, and a universal exchange and inference language in the analysis of socioeconomic health data[☆]

Barry Robson[*], S. Boray

Ingine Inc. Virginia, USA and the Dirac Foundation OxfordShire, UK

ARTICLE INFO

Keywords:
Population health
Socioeconomic
Data analytics
Data mining
Sparse data
Inference net
Hyperbolic Dirac net
Bayes net
Decision support

ABSTRACT

While clinical and biomedical information in digital form has been escalating, it is socioeconomic factors that are important determinants of health on the national and global scale. We show how collective use of data mining and prediction algorithms to analyze socioeconomic population health data can stand beside classical correlation analysis in routine data analysis. The underlying theoretical basis is the Dirac notation and algebra that is a scientific standard but unusual outside of the physical sciences, combined with a theory of expected information first developed for analyzing sparse data but still largely confined to bioinformatics. The latter was important here because the records analyzed (which are for US counties and equivalents, not patients) are very few by contemporary data mining standards. The approach is very unlikely to be familiar to socioeconomic researchers, so the theory and the advantages of our inference nets over the Bayes Net are reviewed here, mostly using socioeconomic examples. While our expertise and focus is in regard to novel analytical methods rather than socioeconomics per se, a significant negative (countertrending) relationship between population health and equity was initially surprising, at least to the present authors. This encouraged deeper exploration including that of the relationship between our data mining methods and traditional Pearson's correlation. The latter is susceptible to giving wrong conclusions if a phenomenon called Simpson's paradox applies, so this is also investigated. Also discussed is that, even for very few records, associative data mining can still demand significant computational resources due to a combinatorial explosion.

1. Introduction and review

1.1. Background

The explosive growth in medical data in digital form has primarily been in regard to clinical and biomedical data [1]. However, this has been primarily of a clinical and scientific nature. The ability of individuals to access and afford the fruits of growing medical knowledge is also a strong determinant of health. Data concerning that is essentially *socioeconomic health data* (SHD). Of course, researchers and organizations have studied data of this general kind for many years. Notably, in pursuit of an understanding of the cause of cholera in the mid-1800s, John Snow had the distribution of cholera deaths and access to national mortality rates by area, and 1849 mortality rates and 8 potential explanatory, primarily socioeconomic, variables for each the 38 registration districts of London [2]. Nonetheless, socioeconomic factors, public perceptions, political opinions and administrative policies can change very rapidly, and modern socioeconomic data relating to public health have been rather sparse, especially in regard to including potential determinants that are now considered as important to modern industrial nations, including fairness to minority or underprivileged groups, as *equity*. From the point of view of our interests in Big Data, this data is small in the sense of comprising just a few hundred records, simply because it addresses counties or county equivalent areas in the US rather than individuals (making it of the order of 100,000 times smaller than potential future patient data in the US case). Such

[☆] This paper is provided to the community to promote the more general applications of the thinking of Professor Paul A. M. Dirac in human and animal medicine in accordance with the charter of The Dirac Foundation, to emphasize the advantages and simplicity of the basic form of the Hyperbolic Dirac Net, to encourage its use, and to propose at least some of the principles of the associated Q-UEL, a universal exchange language for medicine, as a basis for a standard for interoperability. These mathematical and engineering principles are used, amongst many others in an integrated way, in the algorithms and internal architectural features of the BioIngine.com, a distributed system developed by Ingine Inc. VA for the mining of, and inference from, Very Big Data for commercial purposes.

[*] Corresponding author.
E-mail address: barryrobson@ingine.com (B. Robson).

https://doi.org/10.1016/j.compbiomed.2019.103369
Received 10 May 2019; Received in revised form 22 July 2019; Accepted 23 July 2019

data only became publically available in 2018 [3] but is increasing rapidly; the site produced 2019 rankings as early as around March 2019, and hospital admissions and length stay on a county basis also became recently available to us, allowing joining of the data. While making full use of healthcare data overall is still hampered by lack of interoperability and a universally agreed standard for medical records [4], which are issues of great interest to the present authors, it was not a concern in the present study. That is because the data was available in a common basic form of a spreadsheet.

1.2. Purposes of the present paper and the nature of the challenges

1.2.1. Exploring the utility of recent developments in data mining and prediction approaches for analysis of socioeconomic health data

The present paper does not aim to be a medical, healthcare, or socioeconomic research study. Rather, the emphasis is on refining theoretical and technological approaches, exploring matters arising in the data mining of SHD and in drawing inference from it with a proposed universal exchange and inference language discussed below. Traditional statistical approaches were also used in the present study, primarily the long standing technique called Pearson's correlation developed around 1900. This was not least because controversial socioeconomic aspects relating to *equity* were uncovered, unexpected to at least the present authors, also as discussed and reviewed in this paper. Our hypothesis explored here is both (a) that these controversial aspects are real features of the data that should be addressed by government administrators and policy makers, and (b) that modern techniques of data mining and prediction (at least of the kind that we have developed) can support classical statistical methods and verify, further explore, and help characterize, findings of that nature. Note here that while our interest is in developing new tools of broad application, socioeconomic analysis is a well-defined profession, and administrators and policy makers will expect to see confirmation by such qualified professionals using a battery of standard recognized methods. The small size and primarily numerical content of the data means that standard statistical techniques can easily be applied, so possibly such studies are underway by others at the time of writing this present paper. Our relatively unusual methods are based on the use of the Dirac notation and associated algebra from *quantum mechanics* (QM) [5], developed following a speculation that this might provide a basis for basis for a standard best practice in medical decision support [6]. Nonetheless, as discussed below, the techniques of most importance in this present study still belong to the established discipline of structured data mining [7], combined with predictions from the information obtained. Applying both established methods and our newer methods had the benefit of showing how classical Pearson's correlation scores relate to likelihood ratios and predictive capabilities measured in terms of sensitivity, specificity etc. based on association (or associative) data mining. Somewhat surprisingly, we have as yet found no papers *dedicated* to a study of that kind although of course many researchers use both Pearson's correlation and data mining (see Discussion Section 5).

1.2.2. Use of data mining techniques for the relatively small amounts of data from socioeconomic health studies

Socioeconomic data has mostly so far been analyzed by statisticians using classical statistical methods and the data generally provided are of a size and nature that facilitates such analyses. The classical statistical methods developed in the late 1800s and early 1900s are particularly powerful for continuous quantitative data and the continuity in the data values provides significant information that makes use of low amounts of data possible. Even methods that are considered non-parametric in the sense of not using a model such as a normal distribution assume an elementary underlying model. If one has values 10.5 and 15.2 for something, these are not taken as arbitrary names and one assumes information regarding certain statistical properties of values between. Pearson's correlation R also assumes a dominating monotonic trend in the data overall. In contrast, data mining for studying associations between categorical data can easily be applied to numerical data but only by considering it or rendering it categorical, typically by binning into ranges such as 10–19, 20–29 etc. Conclusions can be drawn subsequently about trends by having a continuity model in mind but information dependent on continuity has been lost by a kind of segmentation of the data. This can be a strength when classical statistical models break down concerning the nature of any continuity, as in strongly non-monotonic behavior, or in the potential appearance of Simpsons' paradox considered in this paper. More generally, it also provides fine-grained information about relationships that can be captured and used to make predictions. However it requires data of much greater *depth*, i.e. more records. The data used here comprises only 500 ranked records about the performance of counties or county equivalents, up to some 900 records when distinguished by additional information in additional columns, such as perceived good performance and up-and-coming achievement for each of rural and urban areas (four classes overall). The factors are summary scores for population (community) health with scores for equity, education, economy, housing, food and nutrition, environment, public safety, and community vitality. When joined with annual hospital admissions and length of stay data distinguished by year 2011–2016, there are just some 372 complete records with 25 columns. Very recent 2019 data for 500 ranked records about the performance of counties or county equivalents adds a few more, and we have added to this paper analyses including these, but the data are still sparse for modern data mining. Nonetheless, our less standard data mining approach discussed in this paper is designed to cope with sparse data because data miners, as they drill further down, inevitably encounter sparsity even in Big Data. For example, a probability such as P(A,B,C,D,E,F,G) will have a lot less data to estimate it than would P(A,B,C,D). The joint event (A,B,C,D,E,F,G) may be seen only one or two times, or perhaps not at all, but it does not imply that the probability really is zero as would be seen if a lot more data were available.

1.2.3. The combinatorial explosion challenge persists for the use of data mining techniques, even for small data

Here focus is now on the so-called “width” of the data and on the considerable computer resources that this can often demand. This follows from the final comments in section 1.2.2 above but to researchers not familiar with data mining it can still seem surprising that even data of just a very few records can still provide significant computational challenges, at least *when its fullest benefits are to be obtained by extensive extraction of the information in that data*. The presence of many columns implies a high dimensional multivariate problem aggravated by the cardinality of the data. Such data is said to generate a significant *combinatorial explosion* in terms of the number of possible distinct joint events [1]. That is, sampling exhaustive or probabilistic potentially creates a large number of potentially relevant distinct joint events (A), (A, B), (C, F, H) (G, L, M, Q) etc. For the original population health data used here (prior to joining hospital admissions and stay data), the study had ideally (but not of course in practice) to address potentially some 10^{16} data-mined combinations of 1,2,3,4,5 … non-redundant associated factors at a time, again arising from the combinatorics of columns and the cardinality, the number of distinct value entries in each column. This challenge will become more severe in the future as new data is now being released every year [3]. Of course, the number of counties and county-like entities in the US is not going to change significantly in the near future, but more likely is that more detailed new studies will add new factors, and joining with other county data certainly does. It could increase the number of combinations to beyond astronomic levels (typical current rough estimates put the number of fundamental particles in the visible universe at around 10^{86}). Records of some N = 100 pieces of information would have some 2^N-1 i.e. approximately 10^{30} such combinations even if all entries are binary, and with a more realistic cardinality, 10^{50}-10^{100}. Such numbers could easily be reached by including additional factors, in a future population health study, that plausibly effect health.

1.2.4. The combinatorial explosion problem is in practice ameliorated by the sparseness of data

What reduces the above resource problem to some extent, or at least puts it beyond the responsibility of the researcher, is that when these numbers above arising from the combinatorial explosion are large, then most of the joint events will never be seen. The emphasis therefore shifts to the algorithmic problem of making use of information in sparse data and the relatively few joint events that can be seen (though it may still be a very big number). This is not strategy in traditional statistics: high school teaching states that a study underpowered because of small data should be put on hold until a larger sample is obtained. Unfortunately, in healthcare, long deferral is not always a desirable option, nor is it necessarily appropriate to discard use of data based on the essentially subjective opinion that it is arbitrarily too small to make a useful contribution. When using these results to build inference nets and make decisions, one should take into account that, as in a court of law, many pieces of weak evidence may combine to override a decision that would be made without them. Objectively all available data should be used without arbitrary rejection, and this becomes even more important when, because of limitation in data size, none of the individual counts are large. In addition, one should not discard data solely because it is “poorly supported” in the data, on the basis of rare occurrence. Even a specific item of data about the occurrence of just two or more factors can be highly statistically significant and influence predictions, because the occurrence is *a lot less* than one would expect on a chance basis.

1.3. The Hyperbolic Dirac Net, Q-UEL language, and related work

The Hyperbolic Dirac Net (HDN) [8–14] was developed as a kind of inference net based on Dirac's notation and associated algebra [5,6]. This was extensively used in the present study, although mostly focus will be on odds (ratios of probabilities) in a way that still preserves the essential features of the above. HDN development was followed by the associated Q-UEL language (Quantum Universal Exchange Language) [15–19] formalized primarily in response to a call for such a Universal Exchange Language (UEL) in the 2010 “PCAST report” [4]. While the primary goals of Q-UEL were interoperability and as a proposed standard canonical representation of elements of medical knowledge that would also formalize use of probabilities for evidence based medicine, another major function of Q-UEL was to use those elements as the building blocks for automated and semi-automated HDN construction. Subsequently there was the industrialization of that idea as a high performance system called the BioIngine [20–26].

There is of course a great deal of related work by others (see particularly our Q-UEL/HDN bibliography refs [8–26] that also include review of other efforts), and some provide important pillars for our approach. In hindsight, the HDN and hence indirectly Q-UEL owes a debt not only to Dirac but also to Pearl's Bayes Net (BN) [27], of which the HDN could be considered an extension, not least because any BN could be considered as a subgraph of possible HDNs. Certainly, simplified and crisper rules for building HDNs (while adhering to valid probability laws and without using the same implied information twice or making unwarranted omissions of information) have followed not just from information theoretic considerations discussed later below but also from considering how to transform a BN into an HDN. See Refs. [12,13], and especially ref [14]. Q-UEL also owes a debt to XML [28] and several specific applications of it. The 2010 PCAST report also stated as follows: “We believe that the natural syntax for such a universal exchange language will be some kind of extensible markup language (an XML variant, for example) capable of exchanging data from an unspecified number of (not necessarily harmonized) semantic realms. Such languages are structured as individual data elements, together with metadata that provide an annotation for each data element.” Q-UEL can be considered as an XML extension, although it is also true there is a remarkable fortunate coincidence between the appearance of Dirac notation and XML. This makes Q-UEL a natural convergence of the Dirac notation and an XML specialized for probabilistic semantics. Q-UEL also borrows the RDF model of the Semantic Web to define attributes and relationships in its XML-like tags [29]. Many approaches to representing medical records have been XML-based [18], but to the author's knowledge, only the subsequent Yosemite Manifesto [29], by also being a Web-based approach in response to the PCAST report, came close to Q-UEL in the sense of proposing the use of the Semantic Web and the RDF method.

Despite the above, to our knowledge no alternative approach with similar aims resembles Q-UEL in the sense of having elements that possess algebraic force. This allows Q-UEL's XML-like tags to be used in calculations of probabilistic measures in medicine, and programming of inference nets, inference engines, and expert systems [16,19–26]. This idea is also ultimately also indebted to ideas from other workers in remarkably diverse sources, from mathematics (e.g. see review in Ref. [13]), causal arguments (e.g. Ref. [30] concerning the BN, but see discussion in Section 1.5 (i)). quantum mechanical texts in particular regard to linear operators and dualization (see Theory Section and ref [31]) semantics, linguistics and theories of mind (see Refs. [19–21] for review), and studies of healthcare as a dynamic, complex, and adaptive system (e.g. Ref. [32], and see Section 1.6 below). Sources on socioeconomics and equity [33–43], as discussed in Section 1.6 below, were not only important for the present analysis and its interpretations, but also for promoting the development of some additional support tools related to joining data and novel graphics representations (see Fig. 3 later below). Because the socioeconomic data available was sparse, we relied on data mining techniques for sparse data developed in bioinformatics [44–46].

1.4. Review of the main features of the Hyperbolic Dirac Net that distinguish it from the Bayes Net

We have been invited to review and clarify our approach here with particular reference to its advantages over previous methods. Our approach will certainly be unfamiliar to almost all researchers in socioeconomics. As exemplified in Section 1.5 below, many criticisms can be (and have been) made of the BN, but in our opinion there are few if any proposals that are both (a) definable as being of the same general kind of inference net approach, while at the same time (b) not so close to the BN as to be considered as anything other than valuable BN modifications (see e.g. Ref. [16] for review). Emphasis is on the popular Bayes Net as gold standard. Related to the above, we were also invited to emphasize why the hyperbolic complex (***h***-complex) algebra, and in effect an ***h***-complex Hilbert space, is seen as the best option for overcoming the perceived difficulties of Bayes Nets. We have elsewhere listed a number of advantages of our approach (e.g. Refs. [12–14,16]). The following, however, stand out in particular, for purposes of the present paper.

(i) *Realism.* Perhaps the most general statement of importance is that relationships in probabilistic knowledge used to model the real world are clearly better represented more generally by a Bidirectional General Graph (BGG), essentially what one more typically thinks of as “a network of interactions”. Use of the full BGG is supported by methods described in Refs. [12–14], usually using modules DiracMiner with DiracBuilder [20]. Pearl's popular BN is by definition constrained to represent knowledge by a Directed Acyclic Graph (DAG), popular partly because it simplifies modeling that at least adheres to basic laws of probabilistic relationships and facilitates considerations of causality [27]. But this is not a constraint that we see on interactions in the real world, e.g. in road, train, and subway maps, electric and electronic circuit diagrams, student's mind or concept maps used in study and revision, metabolism, natural neural networks, and not least the complex interplay of physiological systems including feedback for homeostasis that are perturbed in disease. Most notably, these typically include cycles (cyclic pathways) of interactions that the DAG disallows but the BGG allows. Note that in populations sampled for classical statistics, complex multiway interactions, including feedback and other cyclic interactions, occur all the time, and govern the probabilities obtained.

(ii) *Bidirectionality.* More specifically, the essential feature of any model that supports a BGG, and notably cycles, is that it must be able to look in both directions of conditionality at the same time, e.g. P(A|B) concerning how A is dependent probabilistically on B and conversely P(B|A) concerning how B is dependent probabilistically on A. Bidirectionality is an essential feature of the methods discussed above [12–14,16,18]. While the BGG, perhaps including cyclic paths, requires bidirectionality, the converse is not true, since simpler and restricted graph structures can be bidirectional: A ↔ B. Bidirectionaity by itself is important: medicine is not only interested in the probabilities of outcomes, but also in etiologies (causes), and when cause and outcome are understood, study of the conditional probabilities in two directions is important. DiracSmash [26], is an odds-based inference net approach that replaces the BGG by a different concept with similar final effect, but it still supports bidirectional conditionality through the combined use predictive odds and likelihood ratio. See Section 2.3. The BN is in contrast *unidirectional* by definition, and discouraged other considerations at the outset by emphasizing a causal model. It considered that if e.g. A causes B then the converse cannot be true. However, a direct causal relationship between them (one-way or two-way) is not a requirement for a probabilistic relationship, as Bayes rule indicates (Section 2.1), and what determines what should not be presumed, as the well-known adage of classical statistics that "correlation does not imply causation" implies. Finally, the name Bayes Net was chosen to honor Bayes but ironically it cannot contain Bayes' rule, which is inherently about two directions of conditionality [12–14], and which has consequences: See Section 1.5 (i) below.

(iii) *Use of Hyperbolic Complex Probability Amplitudes instead of Probabilities.* While the DAG represents rules for legal construction of the BN, general rules on how to construct a BGG (in which probability laws are also obeyed) were originally less obvious and were tackled on a case by case basis, particularly at branch points and in any cyclic paths. Quantum mechanics (QM) solves the problem by encoding probabilities in both directions of conditionality in a *probability amplitude* that is a single (i.e. scalar) complex number, i.e. with an imaginary part. In the 1930s Dirac developed the now standard physicists' notation to facilitate the use of such algebra by emphasizing the directional aspect [5]. Most commonly in QM one thinks of the imaginary part being multiplied by $\boldsymbol{i}$ for which $\boldsymbol{ii} = -1$, but this ultimately results in a wave description and the notorious weirdness of QM predictions. In contrast, the *hyperbolic imaginary number* $\boldsymbol{h}$ such that $\boldsymbol{hh} = +1$, also rediscovered by Dirac and vital for the theoretical development of QM for high energy particle physics, results in behavior that we expect in the everyday world of human experience (unless $\boldsymbol{i}$ is still present). The structure of QM remains otherwise essentially the same, so that the mathematical machinery succinctly represented by Dirac notation can still be used. The use of $\boldsymbol{h}$ is unfamiliar in medicine, but simple. All that one is ultimately required to know in going from classical Section 2.1. to QM Section 2.2 below is that $\boldsymbol{hh} = +1$. In the authors' opinion, re-expressing the algebra in terms of $\boldsymbol{h}$-complex *spinor projectors* $\boldsymbol{\imath}$ and $\boldsymbol{\imath}^*$ (Theory Section 2.2) makes things simpler still. It is true that $\boldsymbol{i}$ and $\boldsymbol{h}$ can both be replaced by vector-matrix algebra, but it is doubtful that this can be considered simpler. While vector-matrix algebra remains possible and occasionally essential for more elaborate semantic inference involving the action of operators as verbs etc., explicit use of vectors and matrices can mostly be replaced by the elegant symbolic manipulations, essentially an intuitive grammar, that the Dirac notation represents. For a properly constructed estimate *joint* probability, cyclic path, or for certain interesting symmetrical relationships, the value of the $\boldsymbol{h}$-complex imaginary part should be zero. See Theory Section 2.2 This provides an important test because a significant non-zero value signifies failure of coherence due to breakdown of consistency of probabilities required by Bayes' rule [12–14], while also indicating the probability values and direction of conditionality required to restore correct behavior. The $\boldsymbol{h}$-complex algebra quantifies categorical logic and with the Dirac notation provides a probabilistic semantics that maps to natural language [14–19]. Thinking in terms of the $\boldsymbol{h}$-complex valued Hilbert space also leads to some useful new graphic ways of representing evidence based medicine (e.g. Fig. 3 in this paper).

(iv) *Better Estimation.* This follows from the above and is perhaps the most practical and striking consideration, and the easiest to assert: *the most basic kind of HDN, at least, can be considered as a Bayes Net except for the parts of the Bayes Net that are incorrect as unwarranted assumptions that arise from the unwarranted focus on unidirectional conditionality* implied by the DAG [19,21,22]. One consequence of this can put medical probability calculations out by factors of say 2–10 times or worse *for each branch point* in the BN [14]. By any standards, this should be seen as a startling deficiency, but admittedly is it easier to state than to justify, so it is discussed below in section 1.5, and in Theory Section 2.

1.5. Some defenses of the Bayes Net and corresponding counterarguments

Researchers in socioeconomics tend to use "classical" statistical methods and in some instances the BN may seem as unfamiliar, or as contentious, as the HDN. Where interested to explore the merits of inference nets, they may well find the BN to be an attractive or safer choice because of its popularity and some 30 years of use. In fairness to the BN, its supporters argue several benefits, presented and critiqued as follows. These go back to, or at least are related to, Pearl's original points [27], but accounts by supporters may be their own interpretation or modification, especially since the BN is often justified in terms of contentious ideas about causality (discussed first).

(i) The BN traditionally seeks to capture cause-effect relationships and is often said to test or model them. This is also used to justify the use of the DAG. In the present authors' opinion this aspect (usually presented as a positive feature of the BN) is the most objectionable conceptually, although its consequences of needless neglect of interactions arising from the use of the DAG (causal model or not) are of most concern in practice (see points (ii)-(iv) below). Probabilities obtained by observation and counting are statistics governed by the adage in Section 1.4 (ii) above that "correlation does not imply causation". Consequently, many BN users avoid the causal aspect. For the DAG justified in terms of causality, the variables as conditional probabilities in a BN should not be applied in any other order than one that respects order in time, but then rather than treatment of causality being a bonus feature, this would greatly limit the BN's range of application. Even within that range of application, writing one particular partial expansion of a more complicated probability does not imply a unique solution to a causal model. Indeed, many example BNs in the literature that appear to justify causal relations are actually one of many possible *exact* expansions of the joint probability being addressed, so that data were sufficient, the BN is not an estimate and so was not needed to provide an estimate, and as an exact solution was inevitably correct irrespective of the particular expansion used and any causal interpretation. No examples of this have been found concerning socioeconomic data, but consider the long-standing Wikipedia entry on Bayesian Networks [30], P(G, S, R) = P(G |S, R) P(S |R) P(R)) is an exact expansion for any assignment of values G = Grass wet (true/false), S = Sprinkler turned on (true/false), and R = Raining (true/false. The statement following is "The model can answer questions about the presence of a cause given the presence of an effect" [30], but that did not necessarily require the BN and it does require prior knowledge of how rain and sprinklers wet grass and how dry weather causes the sprinkler to be turned on. P(G, S, R) = P(R |G, S) P(S |G) P(G) is equally valid for probabilities based on the contingency table based observed number of occurrence of the events, but rain was not caused by any state of the sprinkler or grass. P(R |G, S) may be a much lower probability than P(G |S, R). However, drawing any deductions about mechanism and process from that requires prior knowledge as common sense and more generally it is a dangerous argument. In complex interactions in, for example, chemistry and metabolism, probabilities as concentrations at equilibria or steady state may be high or low but only kinetic studies, i.e. introducing time, will reveal mechanism and sequence, e.g. that A ↔ C ↔ B, not A ↔ B ↔ C. Somewhat ironically, despite all the above, it seems evident that studying both "observational inference" or outcomes and "causal interventions" or etiologies depends on including interactions in both directions of conditionality. But while the HDN provides this, it does not depend on presumptions about causal mechanisms.

(ii) BNs are considered as very effective by modeling a joint distribution, to "fill the gaps" in the data and probabilities available. An HDN can do this similarly, but by considering bidirectional relationships it has the advantage of an *intrinsic* representation of Bayes' Rule as a basis for consistency between probabilities. A BN certainly does not because Bayes' Rule requires conditionality to be expressed in both directions. Unless specifically present as a guide, Bayes' Rule, especially combined normalization and marginal summation, results in the need for probability values in a way that is counterintuitive to humans (and forms the basis of manty probability puzzles) [12]. This leaves it open to human errors especially whenever there is a subjective component, or otherwise it requires a contingency table of counts or probabilities, without gaps, in order to ensure that probabilities are coherent before applying them to the BN [12]. An HDN user modeling a joint probability (or equivalent in an odds HDN discussed below) and showing that he imaginary part of the ***h***-complex value disappears (see Theory Section 2.2).

(iii) The counterargument to the above would be that, providing care is taken in BN use by preprocessing the data that provides the probabilities, then BNs and HDNs are in general simply different estimates. The "filling the gaps" above relates in part to the fact that inference nets of the BN and HDN kind are at their most useful as estimates of a complicated probability when there is insufficient data for exact calculation of it. If built properly by DAG rules, then the BN is a *valid* estimate. But this is essentially the same issue that was partially addressed in Section 1.4 (iv) above, where it was emphasized that valid estimates are not necessarily good estimates. The DAG-based approach leads to the unnecessary neglect of important interactions for which data would clearly be available, as evidenced by the probabilities that *are* calculated [14]. See also Theory Section 2.1. In short, if constructed correctly, the HDN inevitably must do better at "filling the gaps". Any kind of inference net set up by a researcher could be a poorer model or estimate than it need be, but on the whole the different HDN building modules tend to pick, or allow the user to pick, more complicated probabilities (with as many attributes as possible) that are as free of independency assumptions as the data will support.

(iv) Researchers using BNs have on occasion overcome deficiencies of the BN that arise from omitting interactions, and notably to overcome the extreme assumption that "parent nodes" in the BN are independent (i.e. the events etc. described are treated as if come together at random), because it is confined to the DAG. The main solution is by use of hidden nodes and parameterization of them by iterative methods. However, these are essentially minimization/optimization procedures that are not necessarily trivial, unsuited to large networks especially those with more than one hidden node, can have the usual difficulties such as local entrapment, and are relatively rarely applied [14].

(v) Users can overcome some other major difficulties for a BN that are implied in the above by avoiding estimation of *absolute* probabilities (such as prevalence in populations). Those difficulties evaporate for a BN with use of relative probabilities implied by renormalization, distributions, and use of odds as probability ratios. Obviously, the main difficulty here is that absolute probabilities cannot be calculated. But also, these computations of relative or renormalized values can cause most of the probabilities representing interactions in the net (as well as self or prior probabilities) to vanish as a result of cancellations due to division when they do not directly involve the prediction target [14]. This can sometimes even reduce purely multiplicative inference nets like the BN to a single probability [14].

(vi) In regard to the above aspect, the BN can be argued to be no worse than one of the main forms of the HDN, namely the odds HDN [26], used in the present paper. The odds HDN calculates relative probabilities (i.e. odds) and hence does not calculate an absolute probability. Similarly, it also potentially loses many probabilities, as odds, through cancellation. But in practice, the number of interactions considered is greatly increased, not reduced. The approach is designed to use only probabilities (and hence odds) that contain the target, and this is well suited to HDNs that can contain hundreds of thousands or millions of probability terms on that basis, as exemplified in the present paper, with probabilities that can contain many factors (imply many nodes). It may be argued that this causes it to lose its more diffuse network character in the sense that all nodes connect to the target node, but the quantitative results from features such as branch points and cyclic paths can still be implied in the probabilities used where they can contain many factors. Practical algorithms can be made efficient by using approximations based on expectation measures [14], but whatever the approach they usually still represent estimates that imply more interactions than the typical BN.

(vii) It could consequently be argued that the above odds inference algorithm could equally well be applied to a BN, and that for an odds HDN the bidirectional character, or at least some of its benefits, is lost by discarding bidirectional probabilities, so conferring no relative advantage for the HDN. However, closer inspection proves the latter to be incorrect. Odds as probability ratios in an HDN retain the same essential, including bidirectional, features. The HDN thus remains ***h***-complex, and predictive odds and likelihood ratios now provide the dual. It is still subject to Bayes' Rule, albeit with slight modification due to a cancelation (see Theory Section 2.3 and Eqn. (15)), and benefits arising from that still apply.

(viii) The BN implies a strategy (the DAG) that is relatively easy to describe and for the researcher to use where appropriate, and the notion that something causes something is here particularly helpful. The most obvious criticism of the HDN is that it is mathematically more complicated, and inference net construction is more difficult. Our response should of course be that one must go with the tools that are needed for the task, and that any simplification comes at a price. Nonetheless, BN construction and interpretation are still not trivial, and simplicity and complexity are matters of taste, familiarity, and perception of relevance. Moreover, the HDN technology is largely directed at fully or extensively automatic HDN construction, as an important part of the "AI approach".

1.6. Previous work on the complexity of population health as a system

Prior work of interest to the application domain of the present study has not been reviewed by us, and some essential aspects are as follows. For the relatively small data sample used here [3] "complexity" relates as much to the system itself as it does to the challenges for data analytics. It is the dynamic and often bidirectional nature of interactions within the healthcare ecosystem that can make interpretation difficult. Such complex interactions are particularly evident in the so-called "Meikirch model" promoted by Birchir and Hahn [32]. They see human health and wellbeing as a complex adaptive system (CAS), based on five components. Humans like all biological creatures must satisfactorily respond to (a) the demands of life, so they need (b) a biologically given potential (BGP) and (c) a personally acquired potential (PAP). These are properties of individuals embedded within (d) social and (e) environmental determinants of health. Between these five components of health there are potentially 10 complex two-way interactions. Of considerable current interest [33–36] is how *equity* and *equality* impact, and how they are impacted by, all the potential determinants. Though often thought of as the desire for, and implementation of means to, obtain *equality*, the term "equity" in the data used here appears to resemble equality more closely in terms of the measures on which it is based [3]. However, definitions of the distinction between equity and equality differ. In one commonly held opinion [37] equity involves trying to understand and give people *what they need* to enjoy full, healthy lives, while equality aims to ensure that everyone actually gets the same things in order to enjoy full, healthy lives [37]. In others, equality refers to equal opportunity and the same levels of support for all segments of society, while equity goes a step further and refers offering varying levels of support depending upon need to achieve greater fairness of outcomes [38]. Many definitions converge to the idea that equity is the determinant as recognizing equal rights and making attempts to implement that, while equality is a hoped-for or actual outcome, e.g. "Equity is the means/process. Equality is the outcome or end result of the process" [39]. Some authors appear to switch the two, but whatever their definitions, equity and/or equality are viewed as vital in modern education [40] and social structure [41–43]. The OECD (Organization for Economic Collaboration and Development) appears to rate the US as performing rather poorly compared with a basket of primarily European quite prosperous countries [40] in terms of disparity between the literacy of individuals with and without tertiary educated parents, both for 15 year olds and 26–28 year olds. The disparity seems especially marked for nations when taking account of their prosperity (e.g. as measured by the Legatum Institute [41]. Social philosophers may see it as an aspect of, and further evidence for, *anomie*, a concept introduced by Durkheim in 1893 [42]. Anomie is a condition in which society provides little moral guidance to individuals …" [43]. The natural link here between "guidance" and education is suggestive, since extending the curriculum of an education available to all seems a relatively tractable approach to overcoming any unfairness and misbalance in a social system, at least as a first step. Consequently some attention is given to correlations and associations with education in this paper.

2. Theory and review of the notation and principles

2.1. The classical basis and notation

The underpinnings of the basic method (as well some further aspects by us and others [44–50] used in the present paper), do not replace "classical" data analysis and knowledge representation. However, we argue that they generalize and extend them. Our starting point is the familiar *conditional probability*, a purely real-valued scalar value on the interval 0 … 1, primarily historically introduced by de Moivre. It is also used building block and variable in BNs and in HDNs [12–14]. It is of the following form.

$$P(A\ |B) = P(A, B)/P(B) \tag{1}$$

Traditionally, in "classical" frequentist statistics, it is recognized that such probabilities cannot be seen directly. Eqn. (1) "means in practice" $n(A, B)/n(B)$, where $n(A, B)$ is the number of times A and B are seen, and $n(B)$ the number of times B is seen, if those numbers are sufficiently large. Because the data in this study is very sparse, the contribution of the counts is represented in this study by an information theoretic approach [44–46]. The estimate converges to the above approach as data increases, but it converges to 1 as the data vanishes, then implying zero information $I(A|B) = -\log(P(A|B)$ and appropriately having no effect of a purely multiplicative inference net. See Section 2.3.

Above and throughout this paper, A and B are used in the usual traditional way in theoretical discussions; they represent things, events, states, observation, measurements etc. that can be seen and counted, giving probabilities that represent degree of truth or scope of applicability (or representing an analogous degree of belief). By analogy with parameters in XML, in Q-UEL they are generally called *attributes*, and in medical contexts they are often called factors, e.g. demographic or clinical factors. In a theoretical discussion, by either of A and B we can actually mean several things e.g. one would be a little more explicit by writing $P(A\ |B, C)$ stating three distinct things are involved, or $P(A\ |B, C, D)$ indicating four, and so on. Also, A and B in Eqn. (1) could each be whole logical or other expressions A researcher needs to write, B, C etc. more specifically in practical application. Using our Q-UEL *metadata: = value* notation, one may have for example A as 'type 2 diabetes' and B as BMI(kg/m^2): = 'greater than or equal to 25.0' (here BMI is body mass index). This Q-UEL attribute notation will be self-explanatory, and while a detailed account would reveal further useful features, this is not needed here. Note however that single quotes '' delimit the character strings used as metadata and value when there is ambiguity created by embedded whitespace, which would otherwise signify, by default, the logical operator 'AND'. The vertical bar '|' in Eqn. (1) can be rendered as logical 'IF' or 'conditional upon' or 'given that' or 'from the sample of'.

Most discussions of directional causality in regard to the BN seem to us misleading. In Q-UEL and the HDN, we do allow that a probability P ("A is caused by B") can exist as a probabilistic semantic statement (see later below), but it is not necessarily the same thing as $P(A|B)$. One cannot reasonably say, *in terms of conditional probabilities*, that A is caused by B so that $P(A|B) > 0$ and that $P(B|A) = 0$ because B causing A is impossible. It would contravene Bayes' rule:

$$P(A|B)P(B) = P(B|A)P(A) \tag{2}$$

In any event, having only one of $P(A|B)$ and $P(B|A)$ as zero is of course not what we naturally see in data mining, and importantly a probability can of course have a non-zero probability even if implausible as a causal statement. For example, consider the following.

$$P(\text{'type 2 diabetes'}\ |\ \text{BMI(kg/m^2): = 'greater than or equal to 25.0'},\ \text{Age(years): = '16 to 54'}) = 0.63 \tag{3}$$

$$P(\text{BMI(kg/m^2): = 'greater than or equal to 25.0'},\ \text{Age(years): = '16 to 54'}\ |\ \text{'type 2 diabetes'}) = 0.12 \tag{4}$$

There has been for some time the popular impression that a high BMI causes type 2 diabetes, but there is also recognition that the converse is also true (e.g. see Ref. [12] for discussion). Admittedly, the following would look perfectly reasonable even if interpreted causally.

P(BMI(kg/m^2): = 'greater than or equal to 25.0' | 'type 2 diabetes')

However, Eqn. (3) and (4) also mention an age range. It seems very unlikely that Eqn. (4) can be interpreted as saying that type 2 diabetes also *causes an age range*. In practice, it simply represents the probability of finding a person with that range of BMI and age in a sample of type 2 diabetic patients. Similarly we can study associations in the specifically socioeconomic part of the data repository:

P('Population Health': = 'greater than 75' | 'Economy': = 'greater than 75′, 'Public Safety': = '70′, 'Housing': = '50′, 'Study': = 'US-News2019′) = 0.83) (5)

P('Economy': = 'greater than 75′, 'Public Safety': = '70′, 'Housing': = '50′, 'Study': = 'USNews2019' | 'Population Health': = 'greater than 75′) = 0.072 (6)

It seems much more likely that economy and related factors cause higher quality of population health than *vice versa*, but Eqn. (5) and (6) cannot alone tell us that. The ability to reach the further kinds of probability such as P("A is caused by B") depends on prior knowledge of a kind that we often call common sense, but in many cases the direction of causality is not so obvious *a priori*, as in the above example of type 2 diabetes. But causal or not, the assumption of unidirectional relationship in the BN, being based on the DAG, does lead automatically to an assumption that does have a drastic effect on computation. For example, the Bayes Net (BN) would compute a branch between B and D and E as follows.

P(A ← (B ← D, C ← E)) = P(A |B, C) x P(B |D) x P(C |E) (7)

That is to say, the BN makes the assumption that "parent nodes" of A, namely B and C, are statistically independent of each other, and the specific interaction between B and C is lost in this direction of conditionality. It assumes amongst other things that the following "association constant" value

K(B; C) = P(B, C) / P(B)P(C) (8)

has, in the directional of conditionality that matters computationally, the value 1. This choice affects the estimates of propensities and prevalence including joint probabilities [3,4]. This association value of 1 is hardly even approximately true for most interesting medical data, especially as symptoms, lifestyle, comorbidities, and not least lab results are supposed to provide vital clues to the physician by having strong association values. For example, from our earlier published studies,

K('congestive heart failure'; 'cardiac arrhythmias') = 4.14 (9)

For congestive heart failure with a strongly related comorbidity such as pulmonary circulation disorder it is 6.55 and with renal failure 5.64 [14]. Recall that this issue arises at *every* branch of a BN, and it worsens if there three or more parent nodes in a branch. For medical data, the assumption that an association constant is simply 1 can lead to estimates of joint probabilities by a BN as being in error by a factor of 2–10 times or worse, *per branch point* [14]. Using distributions, renormalized values, or relative probability ameliorates this problem, but it leads to cancelations by division of many of the probabilities, losing even more interactions [14]. Somewhat similar issues to all the above arise when considering a property called *recurrence*, which is invisible to a BN but that needs to be considered for the use of a BGG [14].

2.2. The Dirac braket as a probability dual

QM can handle interactions in both directions and does not have the above problems [31]. This because it encodes both directions of conditionality in a complex number, usually with an imaginary part associated with the imaginary number *i* (such that ***ii*** = −1). We may see, for example,

< momentum(eV-sec):=0.2 | position(Angstrom):=6.3 >

This exemplifies the Dirac braket < A|B > and the encoding of the dual can be expressed as {P(A|B), P(B|A)} as an ***i***-complex *probability amplitude*. However, relating the ***i***-complex case to these component conditional probabilities is fairly complicated, involving Dirac's recipe for observable probabilities using a recipe of "ket normalization" and the "Born rule" [31]. By a Lorentz rotation to the ***h***-complex case (recall, also rediscovered by Dirac) this results in classical probability behavior and a correspondingly simpler treatment. One may write as follows.

< 'systolic BP(mmHg)':=145 | Glucose(mg/dl):=180 >

It has a purely hyperbolic complex value (or purely real if P(A|B) = P(B|A)) that can be expressed as a Hermitian Commutator of empirical adjoint probabilities, e.g. conditional probabilities, and in the equivalent spinor form:

< A | B > = ½ [P(A|B) + P(B|A)] + ***h*** ½ [P(A|B) − P(B|A)] = ιP(A|B) + ι*P(B|A) = {P(A|B), P(B|A)} (10)

Note that the so-called *complex conjugate* of < A|B > , indicated by the asterisk in the manner < A|B > *, changes the sign of the imaginary part, and so < A|B > * = < B|A > and < A|B > = < B|A > *, as is also the case in ***i***-complex QM. The physicist's *spinor projectors are* ι = ½ (1 + ***h***) and ι* = ½ (1-***h***). However, the dual {P(A|B), P(B|A)} follows naturally as a shorthand for ιP(A|B) + ι*P(B|A). It is sufficient to write, for example, the following.

< 'type 2 diabetes' | BMI(kg/m^2): = 'greater than or equal to 25.0′, Age(years): = '16 to 54' > = {0.63, 0.14} (11)

Above, the spinor projectors have the following properties that make proofs and calculations easy: the idempotent property ιι = ι, ι*ι* = ι*, the annihilation property ιι* = ι*ι = 0, and the normalization property ι+ι* = ι*+ι = 1. Applying this with Eqn. (10), it follows that when multiplying such brakets in an HDN, {w, x} x {y, z} = {wy, xz}. Similarly division, addition, subtraction and many other algebraic operations can be independently applied to each component of the dual.

Along with simple extension to the odds as discussed later below, the above provides most of the relevant, less commonly known, algebra that is needed for the greater part of the present paper. However a deeper account provides a greater understanding of some aspects, and is given in the footnote below.[1]

[1] The theory addressed in this paper follows from Dirac's idea of *dualization* and its physical interpretation, e.g. see his discussion following his equation (23) in chapter 2 of ref [31]. There one finds an enigmatic unnumbered equation that can be re-rendered as < A|B > = ι < A|B > + ι* < A|B > , trivially true by the above normalization rule ι + ι* = 1 that implies x = ιx + ι*x, where x is real (or at least not ***h***-complex). But it can be more meaningfully shown to be true from Eqn. (10) and the two other rules of ι (iota) algebra, which also allows z = ιx + ι*y where x ≠ y and z is ***h***-complex. If x and y just happen to take the same value, the result is purely real "by coincidence", but a real value is necessarily the case when x is not a distinct adjoint form of y. This applies for self-probabilities P(A), joint probabilities P(A, B), association constants K(A; B), and odds ratios of evidence based medicine etc. In contrast, conditional probabilities P(A|B) and P(B|A), predictive odds, likelihood ratios (see below) have distinct adjoint forms that imply the existence of a dual and a potential ***h***-complex value. But it is often important that we *can* dualize what is normally thought of as a purely real valued entity; see Section 2.3. In Section 2.5, below, it is also seen how every argument does not need to be interchangeable. **R** in < A | **R** |B > is not, if **R** is Hermitian.

2.3. Tag value attributes: probabilities, odds, and association constants

Values that relate to quantities implied algebraically by the tag (and there may be options as to which are used) are called *tag value attributes*.

< 'Public Safety':='60-69' Pfwd:= 0.5795 | 'Population Health':='80-89' **and** 'Environment':='60-69' Pbwd:=0.0993>

Again note that **and** is often omitted, logical AND being the default, and that a comma ',' is also an alternative (but usually indicates that order might be important for some Q-UEL applications). Note especially the tag value attributes called *Pfwd* (probability forward) and *Pbwd* (probability backward) that must appear on the tags as knowledge elements are stored for future use or are to transmitted via the Internet. If neither is present on the tag nor represented in computer memory (as the value of the braket as a variable) then the default value is 1. In practice, data-mined tags usually contain additional features such as a tag name somewhat analogous to that in XML, and a lot of other detail for web management and provenance. By the above, Q-UEL tags may also carry more tag value attributes than the normal probability dual implies or allows, and applications can optionally use these as alternatives, e.g. odds or association constants instead of conditional probabilities. However, except in special indicated cases (e.g. alternative estimates) they must be consistent with the values implied by the probability dual. For example, an obvious case by eye is that if Pfwd or Pbwd take very low values, certain other attributes of interest must also do so.

Of particular interest are the *assoc* (association constant) and *Ofwd* (odds forward) and *Obwd* (odds backward). Like Pfwd and Pbwd, they too have default 1 if not explicitly mentioned. Ofwd corresponds to *predictive odds*, and its converse or adjoint form Obwd is a *likelihood ratio*.

$$\text{Predictive odds} = PO(A|B) = P(A\ |B)\ /\ P(\textbf{not}\ A\ |B) = P(A, B)\ /\ P(\textbf{not}\ A, B) \tag{12}$$

$$\text{Likelihood ratio} = LR(B|A) = P(B|A)/P(B\ |\ \textbf{not}\ A) \tag{13}$$

The odds dual formed from the above is

$$< A\ |\ B > / < \textbf{not}\ A\ |\ B > \ = \{PO(A|B), LR(B|A)\} \tag{14}$$

It also applies to a whole HDN, i.e. an odds HDN. The following two Q-UEL tag examples illustrate the use of probabilities, odds and association constant for a single positive observation, and also use of defaults. *Pfwd* and *assoc* are omitted in the first case and Ofwd in the second case.

< 'Population Health':='ge50' Ofwd:=2.0368 | **if**:=count:=1 | 'Equity':='ge66' 'County':='Mono County; California' 'Class':='Rural-high-performing' 'Rank':='20' Pbwd:=0.002328 Obwd:=0.0630 >

< Population Health':='lt50' Pfwd:=0.6678 | **if**:=(assoc:=22.2249, count:=1) |'Equity':='ge66' 'Infrastructure':='50' 'Public Safety':='50' 'Environment':='70' Pbwd:=0.07523 Obwd:=32.3214 >

Both odds are useful measures in evidence based medicine and other disciplines. Note that Eqn. (12) gives two forms of predictive odds that are equal because of cancelation of P(B) by division. Related to that, in the case of odds applying Bayes Rule really reduces to meaning the following.

$$\text{Predictive odds} = P(A, B)/P(\textbf{not}\ A, B) = \{\ PO(A|B), LR(B|A)\}\ \{1, PrO(A)\} = \text{likelihood ratio x prior odds} \tag{15}$$

Here the prior odds PrO(A) is defined as P(A)/P(**not** A) = P(A)/(1-P(A))

The *assoc* attribute, being symmetric and applying to the relationship, usually appears in the relationship expression, i.e. between, the two vertical bars '|' so seen as relating to a Dirac operator expression (and an eigenvalue of it). A common construction where it is seen is e.g. | **if:** = assoc: = 6.34) |. Note here the relationship of the above association constant to the braket.

$$< A|B > \ = \{\ P(A), P(B)\ \}\ x\ K(A; B)\} \tag{16}$$

One use of the *assoc* value is that P(A|B), P(B|A) and K(A; B} together allow easy calculation of a full range of probabilities concerned with A and B, plus many important evidence based medicine and epidemiological measures, including odds. Importantly it also provides the basis for the symmetry correction mentioned in Section 2.1 above, and the reomoval of the flaw in the BN that parent nodes at a branch point are independent, one notes first the branch point, e.g. as in Eqn. (5), and then multiplies the net by a corrective form such as the following (required at each branch point)

$$< B, C\ |\ ?\ > / < B|? > \ < C|? > \ = \{K(A; B), 1\} \tag{17}$$

See also ref [14] for a rather similar correction in the other direction of conditionality, relating to the idea of occurrence of things, events, states etc.

Note above the special attribute '?'. Here '?' is the event of preparation or observation that certainly occurred, even if what is set or measured is uncertain, so that P(?) = 1. Terms like < A|? > and < B|? > or < ?|A > and < ?|B > are also needed as analogues of simple, self, or prior probabilities P(A) and P(B), but which can now be distinguished in two directions of conditionality, and they are most generally ***h***-complex. Both the HDN and the traditional BN typically use self or simple, as prior probabilities, updated by the rest of the net. In Q-UEL and the HDN one defines them as follows.

$$< A|? > \ = \boldsymbol{\iota} P(A)\ +\ \boldsymbol{\iota}^*\ =\ \{P(A), 1\} \tag{18}$$

$$< ?|B > \ =\ \boldsymbol{\iota}\ +\ \boldsymbol{\iota}^*\ P(?|B)\ =\ \{1, P(B)\} \tag{19}$$

2.4. Estimation of probabilities and brakets from sparse data

The Q-UEL approach also has roots in an information-theoretic strategy originally developed in bioinformatics for sparse data [44–46]. A Baeysian degree of belief can be incorporated: the n[ ] are frequencies of observation (numbers of occurrences), which could be zero, but the v[ ] are virtual frequencies reflecting prior belief, e.g. such that belief Be (A|B) = v[A, B]/v[B]), calculating v[A, B] = Be(A, B) $_x$ N for total data N, and so on. In the present study all v[ ] = 0. These "expected frequencies" contribute to, but are distinct from, the expected conditional information I(A|B,C, …) given data d[A,B,C, …]. It gives an *expected probability* P(A|B,C, …) computed as follows.

$$P(A|B,C,\ \dots) = e^{E(I(A|B,\ \dots)\ |d[A,B,\ \dots])} = e^{\zeta(s=1,\ n[A,B,C,\ \dots])\text{-}\zeta(s=1,\ n[B,C,\ \dots])} \tag{20}$$

In the above equations, ζ(s, n) is the partially summated Riemann zeta function:

$$\zeta(s, n) = 1 + 2^{-s} + 3^{-s} + \ldots. + n^{-s} \tag{21}$$

Note that ζ(s,0) = 0. For s = 1 it emerges naturally from an integral over possible information values each with a Bayesian posterior degree of belief [44,45]. In the present study, a modified form called the z function [20] is used to take account of a small discrepancy between natural logarithms and zeta functions [20]. The distinction is smaller when one zeta function is subtracted from another as here.

In practice, data mining in the BioIngine usually delivers the following value for a braket in spinor form

$$< A \mid B > = \iota\ e^{\zeta(s=1,\ n[A,\ B]+v[A,B])-\zeta(s=1,\ n[A]+v[A])} + \iota^{*}e^{\zeta(s=1,\ n[A,\ B]+v[A,B])-\zeta(s=1,\ n[B]+v[B])} \tag{22}$$

As noted above, Q-UEL tags ultimately intended for general use and sharing carry a great deal of detail for management on the World Wide Web, as well annotation as to provenance. Provenance includes details pertaining to use of zeta functions and comparison with tradition calculations e.g.

<Q-UEL-DIRACMINER-KMETHOD-3-FACTOR-POPHEALTH-SURVEY:=(application:='Perl version v5.16.3':=DiracMiner158.txt, input:=CommunityHealth.csv, patient/population#:=all:=0-900, samplesize:=900, incidences:=22, prior=0, tagtime(gmt):='Sat Dec 8 13:24:10 2018')

'Environment':='60-69' Pfwd:=(Pfzeta:=0.4702:=exp[3.1387-3.8934], classical:=0.4583:=22/48)

| **if**:='do all':=(assoc:=(standard:=1.4373, atomic:=(strength(nats):=3.1499:=3.1387--0.0112, Kzeta:=1.2784:=exp[3.1387+13.6048-5.5016-5.3331-5.6632], classical:=1.2353:=22*8/244*206*287)), Pjoint:=0.0244:=22/900), events:=(P[Education:=40-49]=0.2723~=244/900), P[Housing:=60-69]=0.2301~=206/900), P[Environment:=60-69]=0.3201~=287/900)) |

'Education':='40-49' and 'Housing':='60-69' Pbwd:=(Pbzeta:=0.0801:=exp[3.1387-5.6632], classical:=0.0767:=22/287)

Q-UEL-DIRACMINER-KMETHOD-3-FACTOR-POPHEALTH-SURVEY>

When all virtual frequencies v[ ] are zero, the above kinds of Q-UEL tag value conform with Q-UEL's use of default 1 for Pfwd etc. Because the probabilities approach 1 as the total amount of data N approaches zero. However, we rapidly approach classical probabilities, usually around N = 20, as N increases. It reflects zero information theory which states that P(A) = 1 if information I(A) = -log(P(A)) = 0. This is held to be consistent with *Popper's principle of evidence and refutation* [47]. A statement unqualified by probability is *either* a statement tentatively thought of as having probability 1 awaiting refutation by contrary evidence, *or* is an assertion for which there is lack of knowledge or hard evidence (which may be the same thing). In most real world applications, the user of an inference net never has access to the vast number of influences that may impact the calculation. Not including a tag or its equivalent in a purely multiplicative HDN (or a BN) because we lack data, knowledge or even awareness, or discarding those where statement of knowledge that it implies makes no sense, is the same as including it with probability 1. The overall benefit of this approach to sparse data is in the HDN inference net context, where as in a court of law a lot of weak evidence can be brought together to overthrow a decision made without it.

2.5. Semantic information and automated hunt for information on the internet

In the present study, the following relates to literature research to support studies carried somewhat behind the scenes rather than direct used in inference [19,21,22], but it is part of our BioIngine methodology, and we seek to raise it to represent a fundamental aspect of the automatic or semi-automatic decision making process. It constitutes a kind of semi-automated systematic review [23], and follows from the same Q-UEL basic theoretical principles. Dirac's *bra-operator-ket form*, where **R** can represent a relationship (or relationship operator or *relator*) such as a verb, giving < A | **R** | B > . In general, we can more explicitly write.

< subject expression | **relationship expression** | object expression >

The operator **R**, or relationship expression generally, is in all interesting cases Hermitian (as is so in QM), such that

$$< A \mid \mathbf{R} \mid B > = < B \mid \mathbf{R}^{*} \mid A > = < B \mid \mathbf{R} \mid A >^{*} \tag{23}$$

Here < B | **R*** | A > the active/passive inverse of < A | **R** | B > , but complex conjugation < B | **R** | A > * meaning (< B | **R** | A >)* switches subject and object. If it wasn't, then one would be writing e.g. < A | **R** | B > = < B | **S** | A > *. For example, < A | **if** | B > is often used for < A|B > to present the semantic triple form, but one is now free to make semantic interpretations of < A|B > such as < B | **are** |A > (categorical interpretation) or < B| **causes** |A > = < A | '**is caused by**' B > (causal interpretation), though the latter interpretation in particular still requires external evidence. It is often the case it equals < A|B > , and obviously so if a causal interpretation of < A|B > is taken, but not in general. Except for some special Hermitian matrices, Hermitian character of the relationships requires that < A| and |B > are orthogonal vectors, which in logical and semantic terms means that A and B are mutually exclusive. Then < A|B > = < B|A > = 0. This is however no worse than an interpretation such as the homely example < dogs | **if** | cats > = < cats| **are** | dogs > = 0, while < dogs | ***chase*** | cats > is greater than zero (and the inverse is too, but with a lower probability). Note that we can perform logical, relational, and syllogistic reasoning, as with the simple example < A | **are** |C > = < A | **are** |B > < B | **are** |C > . in contrast, a construct < dogs | ***chase*** | cats > < cats | ***chase*** | mice > has a numerical value that carries little more information than stating collective truth of the component statements, of the scenario as a whole, though there can be further utility if there is for example an ecological or epidemiological effect of such a chain of events, essentially impactful or even causal, e.g. propagating disease by fleas [19]. Notably, the researcher should somehow be able to see or imagine A causing B, as in seeing dogs chasing cats, and in principle these are also events that can be observed, sampled, and counted.

Although form < A| **if** | B > is the most prevalent tag form used in this paper, in order to use the form of the semantic triple (subject-relationship-object), more interesting verbal relationships did arise in support studies in the present project. A Q-UEL application XTRACTOR searched Internet text to obtain the statements of the above kind, parsing extracts of text, usually sentences, and putting them in a format from which that from which these forms < A | **R** | B > , which may be may be thought of as "semantic triples", or forms like < A |**R** |B |**S** |C > = < A | **R** | B > < B | **S** | C > as "linear semantic multiples" are readily generated. In effect, XTRACT tags "autosurf and spawn" on the

Internet. Links and citations in Q-UEL tags illustrated below point XTRACTOR to go to an process a new web page, generating still more such XTRACT tags, and this continues in an explosive process. In XTRACT tags, annotation [0htpp …] signifies a reference that was an in-text link. Not appearing in these examples is that references found at the foot of the web page appear as [1http...], [2http …,] etc. [16,22].

```
<Q-UEL-XTRACT "Population _health |^has ^been ^defined as| `the health _outcomes |of| `a _group
|of| individuals (including) |as| `the _distribution |of| `such  outcomes |within| `the group&#911&#93;
(?_it) Population _health |^is| `an approach |to| health [0https://en.wikipedia.org/wiki/Health] |^aims to
^improve| `the _health [0https://en.wikipedia.org/wiki/Health] |of| `an `entire human _population; (This?)
human _population (concept)|^does not ^refer to| _animal {OR} plant populations; (This?) human
_population (concept)  |^has ^been ^described as ^consisting of| `three components"
| 'was extracted from' |
source:='https://en.wikipedia.org/wiki/Population health' time:='Thu Dec 13 15:48:13 2018' extract:=64 Q-
UEL-XTRACT>

<Q-UEL-XTRACT "`(?These) |^are| health_outcomes {AND} patterns |of| health {AND}  policies {AND}
interventions [0https://en.wikipedia.org/wiki/Population_health_policies_and_interventions]&#911&#93
|as| `A _priority |^considered in ^achieving ^aim of| Population _Health |^is to ^reduce| health
inequities {OR}  disparities |among| different population _groups |among| `other _factors `the social
_determinants |of| health [0https://en.wikipedia.org/wiki/Social_determinants_of_health] |as| SDOH"
| 'was extracted from' |
source:='https://en.wikipedia.org/wiki/Population health' time:='Thu Dec 13 15:48:13 2018' extract:=65 Q-
UEL-XTRACT>
```

The slightly unnatural reading of some of the above sentences is due to the fact that it is not simply extracted text. Parsing occurs followed by reorganization of sentences (or combinations or parts of them) such that the sentence structure, in general a tree graph, is recomposed as much as possible into linear semantic multiples as above, that facilitate extraction of semantic triples $< \mathrm{A}|\ \mathbf{R}\ |\mathrm{B} >$. XTRACTOR tries to make interpretations that may need curation, e.g. concerning what pronouns, correlatives etc. actually refer to (e.g. what is referred to by “it”). The reason for the additional symbols (such as prefix ^ for verbs) and added annotation is to disclose the parsing and interpretative assumptions that were made, because sometimes correct interpretation fails and curation (by humans or more automatically) is required.

3. Methods

3.1. Data

The original data [3] comprised one record for each of 500 counties or similar regions that are top scorers in population health, additionally split into 4 classes of 100 that represent urban high performing, urban up-and-coming, rural high performing, and rural up-and-coming. Along the above grouping (urban high performing, etc. plus the class Overall) the data comprises only scores for population health with 9 other types of score such as Economy, Infrastructure, and so on each of which already represents a curation and pooling by combining scores from several sources. The later 2019 data as used by us (see below) also included a score ‘Healthiest Community Score’ as well as the important ‘Population Health’; the former is not specifically medical. To this we add the County or county equivalent, each of the above 4 classes and the County and in some studies a preliminary computed normality score as to how usual, or otherwise, the record is. That makes 11 columns in all, but still a relatively small number. Although the many data mining examples used the 2018 data that was that available [3] when accessed on 12/16/2018, extensive comparisons have been made with recent 2019 data. The 2018 data or overall ranking for the top 500 ranks counties etc., corresponding to class ‘Overall’, is almost always used combined with the rankings for the top counties etc. ranked in the top 100 of each of the urban high performing, urban up-and-coming etc., comprising 900 records overall. These 900 records had the considerable advantage over the basic 500 in that they included a wider range of values particularly for Population Health which would otherwise be almost entirely confined to the range 50–100; this combination caused some counties to be counted twice and this was retained in order to provide more data-minable information, i.e. with the records distinguished by the above classes and their rank in them. Notice that when referring to factors as indicated by the name defined as the metadata (column headers), we generally use throughout capitals for brevity, e.g. “Population Health” as opposed to writing e.g. “population

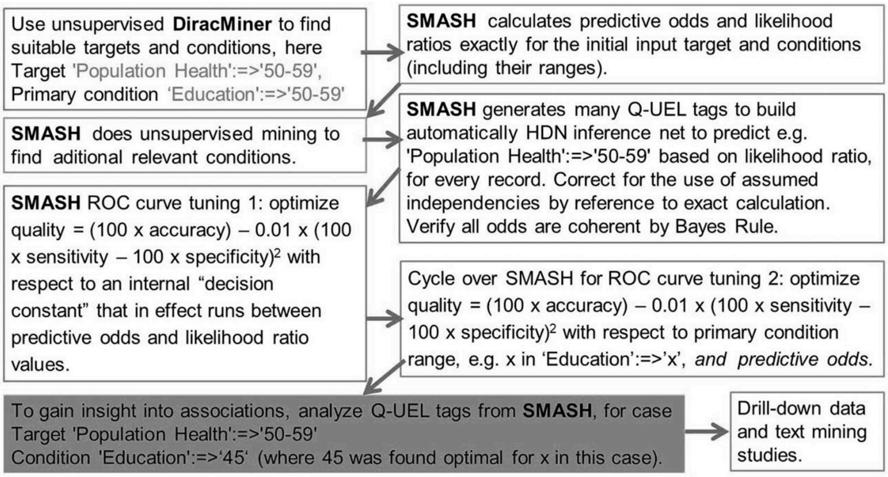

Use unsupervised DiracMiner to find suitable targets and conditions, here Target 'Population Health':=>'50-59', Primary condition 'Education':=>'50-59'
SMASH calculates predictive odds and likelihood ratios exactly for the initial input target and conditions (including their ranges).
SMASH does unsupervised mining to find aditional relevant conditions.
SMASH generates many Q-UEL tags to build automatically HDN inference net to predict e.g. 'Population Health':=>'50-59' based on likelihood ratio, for every record. Correct for the use of assumed independencies by reference to exact calculation. Verify all odds are coherent by Bayes Rule.
SMASH ROC curve tuning 1: optimize quality = (100 x accuracy) – 0.01 x (100 x sensitivity – 100 x specificity)² with respect to an internal “decision constant” that in effect runs between predictive odds and likelihood ratio values.
Cycle over SMASH for ROC curve tuning 2: optimize quality = (100 x accuracy) – 0.01 x (100 x sensitivity – 100 x specificity)² with respect to primary condition range, e.g. x in 'Education':=>'x', and predictive odds.
To gain insight into associations, analyze Q-UEL tags from SMASH, for case Target 'Population Health':=>'50-59' Condition 'Education':=>'45' (where 45 was found optimal for x in this case).
Drill-down data and text mining studies.

Fig. 1. Tools and workflow for the larger part of the present study.

health score", though in the present series of studies capitalized metadata also distinguishes the original data to which other data is joined. The 2019 data is essentially analogous to the above 500 records in Class 'Overall' and was used in the analysis of length of hospital stay. Hence the metadata (here meaning column headers) became as follows.

Healthiest Community 2019 Rank, Community, Healthiest Community Score, Community Vitality, Equity, Economy, Education, Environment, Food & Nutrition, Population Health, Housing, Public Safety, Infrastructure, number admissions 2011, number admissions 2012, number admissions 2013, number admissions 2014, number admissions 2015, number admissions 2016, average length of stay 2011, average length of stay 2012, average length of stay 2013, average length of stay 2014, average length of stay 2015, average length of stay 2016.

The joining of the 2019 and hospital data with the hospital data took account of state and county codes to ensure that counties and county-like entities were not wrongly assigned (some county names are popular and occur in several states), and to ensure that they did not occur twice in this data even with correct assignments to states. Consequently, this data is purer for statistical analysis on a county basis in that each county only appears once in the data set. However, only 372 counties were common to the two sets of records joined as available to us at this time, so resulting in 372 joined records out of the original 500 socioeconomic health records for 2019. Note that values under Community become for example 'Douglas County; Colorado', i.e. they include the state, again important because some popular names are found in many states.

3.2. Tools and Workflows

Although Q-UEL has been defined as a vector-matrix and semantic programming language (16, 19, 23), few Q-UEL applications have as yet been encoded in Q-UEL itself. The *developmental* suite of Q-UEL applications has over some years been written in the Perl language with some past experiments in Python and Mathematica and Wolfram languages [23]. In the BioIngine as the industrialized version, the applications are currently recoded in the Scala language, well suited for scalability in a middleware environment specialized to handle "Very Big Data". The applications may also produce comma separate value (CSV) output files to invoke standard statistical methods available in Python and Microsoft spreadsheets.. The Q-UEL software applications used here were mainly DiracMiner with DiracBuilder 20] that convert CSV files to Q-UEL (and to some natural language text as reports to the physician etc.). There was also use of text analytics by MARPLE/HDNstudent that converts natural language text to Q-UEL, [22,23], and DiracSmash in normal and normality score mode, and ALERT and BILL [26]. The latter three all produce Q-UEL tags as important output, but also convert CSV files to reordered and/or selected subsets of original records (in DiracSmash normality score mode with scores added to records), again reappearing as CSV files.

Refs [22–26] include an account of performance in a more clinical context, and Fig. 1 captures the main use of DiracMiner and DiracSmash in this paper. DiracSmash is interesting because it uses the following as a simple guideline to constructing very large odds-based inference nets, and also in some sense represents the output as simplified prediction model developed interactively. It normally requires a *target* such as 'Population Health': = 'ge50', and a *Hitlist* a list of suspected *interdependent* factors as conditions or denominators, typically demographic. It is a logical AND list. Optionally and commonly, there is also a *Wishlist* or "*Shortlist*" that represents suspected *independent* or *individually sufficient* factors, though they can "add up" to have a stronger effect. It is a logical OR list from the perspective of selecting tags to be used in the HDN and in testing simplified models, but RAND (randomly associated AND) between odds in the HDN; that is they represent the independencies that are the assumptions characteristic of an inference net. In clinical contexts, they are often symptoms and clinical measurements. As indicated by the example tags used throughout this paper, binning of scores prior to sampling was automatically applied in the ranges of 10 such as 50–59, 60–69, 70–79, 80–89, 90–99. Below 10 binning was to ranges of 1, above 100 to ranges of 100, and so on, but these did not arise in the current data except for the rank in each class such as urban-up-and-coming. However, ranges set as e.g. Equity ≥66 are set prior to this binning, and correlation is done prior to binning. For the present studies, the above binning was essential in both DiracSmash and DiracMiner. When DiracMiner was used, data mining was unsupervised. When DiracSmash was used, Wishlist was for the most part not required because focus was on predicting population health scores above a specified value given a single condition such as the equity or economy score above or below a specified value. Extensive high dimensional data mining by DiracSmash was therefore usually supervised only by the Target and Hitlist entry as queries, which enabled easy comparison with Pearson's two-way correlation. This still allowed large inference nets to be built to confirm a positive or negative prediction (usually for a high population health score), for which DiracSmash confirmed coherence (as consistency of probabilities by Bayes' Rule, and correct normalization and marginal summation). The above positive or negative result was used along with the Hitlist condition, for optimizing sensitivity and specificity, and generating simplified prediction models, as follows.

3.3. ROC curve tuning

A strong feature of DiracSmash is that it not only makes predictions but helps propose simplified models which can be considered as "risk scores" for some specified target. The simplest model, as suggested by the initial phase present studies and then used throughout, comprised the Target, a single Hitlst condition, the positive ("yes") or negative ("no") nature of the prediction of the Target given the Hitlist condition, and the optimal value of the threshold as obtained by ROC curve tuning [49]. Note that the likelihood ratio drives the prediction model in DiracSmash and a likelihood ratio of less than 1 causes the prediction to flip automatically to the complementary form. For example, the target Equity ≥66″ becomes "Equity < 66". Roughly speaking, the optimization by the ROC curve [49] as applied in DiracSmash [26] in effect, albeit indirectly, moves the threshold (decision point) between likelihood ratio as predictor and predictive odds as predictor. To obtain reasonable balance between sensitivity and specificity, the threshold is adjusted to maximize the following.

$$\text{Quality} = (100 \text{ x accuracy}) - 0.01 \text{ x } (100 \text{ x sensitivity} - 100 \text{ x specificity})^2 \quad (24)$$

The ROC Match as 'yes' or 'no', and hence as a true or false positive and true or false negative respectively, usually depends at least in part on the Hitlist and Wishlist and the number of factors in them with a different weight. In this study WishlistWeight = 1 so that it dependent wholly on Hitlist and Wishlist with an equal weight. Each optional algorithm has different adjustable weights for the various influences. In the following algorithm as used here, the DirectionalInfo usually corresponds to the exponential of the log-odds-like sum of the information (positive for. and negative against) the target contributed independently by each of all the factors on Hitlist and Wishlist.

$$\text{fractionHits} = (\text{countHitlistHits}+1+(\text{countWishlistHits}+1) \text{ x WishlistWeight}) \,/\, (\text{numberOfHitlistEntries} +1+(\text{numberOfWishlistEntries} +1) \text{ x WishlistWeight}) \quad (25)$$

Table 1
Preliminary studies with the 2018 data. Classic Correlations between potential determinants of Population Health in comparison with Predictive Capability based on Associative Data Mining. The data combined the five classes urban high performing, urban up-and-coming, rural high performing, and rural up-and-coming, and overall.

Target for prediction is Population Health ≥87 (except average stay) predicted with the following as fixed condition. (ROC curve tuned with respect to range threshold – see text)	Classical Pearson correlation of Population health 2018 in same data with 5 classes used for HDN construction. That just for class 'Overall' is shown in brackets.	Size of HDN (number of Q-UEL odds tags). Original Population health data had 13 columns, 25 when joined with hospital stay.	Likelihood ratio computed from HDN. If < 1 then prediction is for Population Health < 50	Predictive capability of HDN predicting Population Health > 50 or < 50 has this sensitivity%	Predictive capability of HDN predicting Population Health > 50 or < 50 has this specificity%
Economy ≥ 51	+0.45,(+0.15)	197,286	**20.733**	80	91
Infrastructure ≥ 60	+0.40,(+0.28)	185,444	**4.693**	72	81
Education ≥ 45	+0.38,(+0.18)	184,168	**3.109**	72	74
Food & Nutrition ≥ 61	+0.33,(+0.24)	144,881	**2.261**	52	74
Public Safety ≥ 67	+0.21,(+0.10)	158,905	**1.856**	57	67
Comm. Vitality ≥ 60	+0.15,(+0.04)	165,987	**1.586**	61	60
Housing ≥ 59	0.00,(-0.01)	118,679	0.692	*57*	*60*
Environment ≥ 65	+0.07,(+0.04)	109,962	0.571	*65*	*60*
Equity ≥ 66	(**-0.33**),-0.32	116,450	**0.534**	61	70
Equity < 66	(+ **0.33**), + *0.32*	168,853	**1.982**	61	67
Average days hospital stay 2016 < 4.5	−0.13	7,646,997 (25 columns)	**0.980**	65	49

match = 'yes' if fractionHits > = matchThreshold & directionalInfoWeight x directionalInfo ≥ matchThreshold (26)

However, for most studies the Wishlist was empty, as if wishlistWeight = 0. Also, initially we used directionalInfoWeight = 10, and given the values of information prevailing in this study, the directional information served primarily as a break on strong contrary evidence. In practice, removing this influence by setting directionalInfoWeight = 0 produced almost identical results. In all the studies, the essential model was almost entirely based on the likelihood ratio predicted by the HDN (as 1 or more, or less than 1) and the minimum fraction of Hitlist and Wishlist entries (factors) that would have to be on a record for it to register as a positive prediction of the target. For ROC curve tuning, this prediction was then compared with the actual value on the target to record it as a true positive or a true negative, or a false positive or a false negative, from which sensitivity, specificity, and other performance metrics are calculated. A very low threshold value for this essentially means that just one determinant as condition is sufficient to generate a positive prediction, and since there is just one such determinant in many of the present studies, this is typically the case. Normally, increasing the threshold value further will cause the quality measure to deteriorate, so that there is a clear optimum, although for most studies here, as described below, an highest value of Quality was obtained early at threshold 0.05 (reported as 5%) and persisted as a plateau.

3.4. *Detection of anomalous records as unusual data points*

For such sparse data as applies in the present study, it is possible that a rare "outlier" case of extremely good or very bad population health could depend on a one-off factor or combination of factors that could be important in providing a clue on how to improve population health. Important supporting roles were also played by three approaches to anomalous record detection, pattern discovery, techniques originally applied to give alerts to patients at risk, and text analytics. DiracSmash in Normality score Mode, ALERT and BILL were all employed [26]. ALERT essentially compares every record with every other record (or a sample selection of other records) with an adjustable model for similarity/non-similarity. BILL also looks for records that represent discrepancies but focuses on a single critical parameter such as cost of an insurance claim value or, in this case, Population Health, assuming that this value is at least roughly proportional to a sum or average of all the other factors (scores). It learns the values of various items on records in order to predict some such total cost or benefit, noting cases when that differs markedly from what is actually claimed to be the value on a record. BILL performed somewhat poorly suggesting that such simple additivity applies only to a very weak extent.

4. Results

4.1. *Preliminary studies and overview of principal statistical results*

Although this paper is primarily methodological (and not medical or socioeconomic research) it is instructive to follow, with some discussion, a story for real data in a practical use-case context. A very early step in most studies is to generate a large Q-UEL tag (a collection of summer data for each specified study) called the SSMETHOD (statistical summary) tag. The following tag labeled POPHEALTH-SURVEY-SUMMARY was generated by DiracMiner in its original Perl version [20] which carries the "standard statistical summary report" usually applied at the beginning of each study, not least to report on the composition of the data. For the data of Table 1 below but confined to the 500 records of class 'Overall', which is the overall ranking for the top 500 ranks counties and county equivalents. The tag shown in part it is as follows.

Table 2
2019 Classic Correlations between Potential determinants of Population Health and length of hospital stay in comparison with Predictive Capability based on Associative Data Mining. This data is class 'Overall' only, and does not include classes such as urban high performing.

Target for prediction is Population Health ≥87 (except average stay) predicted with the following as fixed condition. (ROC curve tuned with respect to range threshold – see text)	Classical Pearson correlation of Population health in the class "Overall" for factors in previous column. The 2018 data is compared in the brackets, the 2019 data follows.	Size of HDN (number of Q-UEL odds tags). Original Population health data had 13 columns, 25 when joined with hospital stay.	Likelihood ratio computed from HDN. If < 1 then prediction is for Population Health < 87	Predictive capability of HDN predicting Population Health > 87 or < 87 has this sensitivity%	Predictive capability of HDN predicting Population Health > 87 or < 87 50 has this has this specificity%
Food & Nutrition ≥ 60	(+0.24), +0.30	89,104	**2.7528**	75	77
Education ≥ 62	(+0.18), +0.26	155,904	**2.660**	65	75
Infrastructure ≥ 77	(+0.28), +0.36	174,046	**2.571**	71	72
Economy ≥ 66	(+0.15), +0.26	207,968	**2.317**	80	65
Public Safety ≥ 72	(+0.10), +0.07	158,905	**1.856**	57	67
Housing ≥ 59	(-0.01),-0.19	212,402	1.302	*69*	*59*
Environment ≥ 69	(+0.07), +0.12	222,398	1.264	*65*	*55*
Comm. Vitality ≥ 60	(+0.04),-0.04	245,780	**1.072**	54	49
Equity ≥ 56	**(-0.32), -0.445**	275,194	**0.288**	77	71
Equity < 56	***(+0.32), +0.445***	168,853	**2.748**	77	72
Target is Population Health ≥ 76 given average days hospital stay 2016 ≥ 5.0 given.	−0.13	1,226,742 (25columns)	**1.197**	51	56

Table 3
Some computer "experiments" in correlations with length of hospital stay in comparison with Predictive Capability based on Associative Data Mining, in order to demonstrate some behavior with respect to the range value.

Target for prediction and "given factor". (ROC curve tuned with respect to range threshold – see text). Unless otherwise specified, the results correspond to the joined data.	Classical Pearson correlation of Population health in the class "Overall" for factors in previous column. The 2018 data is compared in the brackets, the 2019 data follows.	Size of HDN (number of Q-UEL odds tags). Original Population health data had 13 columns, 25 when joined with hospital stay.	Likelihood ratio computed from HDN. If < 1 then prediction is for Population Health < 87	Predictive capability of HDN predicting Population Health > 87 or < 87 has this sensitivity%	Predictive capability of HDN predicting Population Health > 87 or < 87 50 has this has this specificity%
Target is Population Health ≥ 76 given average days hospital stay 2016 ≥ 5.0 given.	−0.13	1,226,742 (25 columns)	**1.197**	51	56
Target is average days hospital stay 2016 ≥ 5.0 given number of admissions 2011 ≥ 4,000, full 2173 admission records only (not confined to highly ranked counties or equivalents).	+0. 115	2,224,246 (15 columns)	**1.005**	49	51
Target is average days hospital stay 2016 ≥ 5.0 given number of admissions 2011 ≥ 6500.	+0.295	2,247,714 (25 columns)	**1.707**	62	62
Target is average days hospital stay 2011 ≥ 5.0 given number of admissions 2016 ≥ 6500.	+0.23	2,260,538 (25 columns)	**1.843**	67	62
Target is average days hospital stay 2011 ≥ 5.0 given number of admissions 2011 ≥ 6500.	+0.295	2,223,785 (25 columns)	**1.856**	67	62
Target is average days hospital stay 2016 ≥ 5.0 given number of admissions 2016 ≥ 6500.	+0.295	2,275,086 (25 columns)	**1.695**	62	61

```
<Q-UEL-POPHEALTH-SURVEY-SUMMARY:=(application:=' v5.16.3':=DiracMiner158.txt,
tagtime(gmt):='Tue Jan  8 02:48:17 2019':='standard cardiovascular report':='relator after clinical
descriptor':=(name:='Rank', number:=1))
patients:='input file':=AetnaOverallClass.csv:=('original sample size':=500, 'selected sample
size':=500, 'number of descriptors':=10)
stakeholder:=initiator:=person:=(biostatistician, developer):=name:='Barry Robson'
stakeholder:=initiator:=organization:='data collection':=http://caymanheartfund.com/contact-us/',
stakeholder:=system:=initiating file:='command file':=Qcommand.txt:='containing initiating
tag':='(rsvdchar open bracket)Q-UEL-POPHEALTH-SURVEY-REQUEST patients:='input
file':=AetnaOverallClass.csv columns:=0-1 (rsvdchar bar) have:=if:='do all' (rsvdchar bar) Q-
UEL-QUERYTAG-POPHEALTH-SUMMARY-REQUEST(rsvdchar close bracket)'
stakeholder:=system:=application:='Perl':=version:='Perl version v5.16.3':='application
name':=DiracMiner158.txt
stakeholder:=system:='input data file':='AetnaOverallClass.csv'
stakeholder:=system:='output data file':='output.txt:=comment:='contained this tag Q-UEL-
POPHEALTH-SURVEY. Standard redirection if not disposed to screen.'
comment:=('***STATISTICAL NOTATION USED BY GENERATING APPLICATION NAMED
DiracMiner158.txt***',
'Reporting (mean value)+/-(dispersion), i.e. meaning range mean-
dispersion...mean+dispersion.',
'Dispersions are P95 (range holds 95% of values), SD standard deviation, SE standard error, CI
confidence interval.')

 Rank:='250.50+/-(282.90P95, 144.34SD, 144.92SDs, 6.45SE, 12.65CI, 0.02Skew, -
2.94Kurtosis)'

Pfwd:='assuming random association':=183.05%:='-logPfwd':=-2.102872

| have:=if:='do all(do all)' |

 County:=('Falls Church city; Virginia':=0.20%, 'Douglas County; Colorado':=0.20%, 'Los Alamos
County; New Mexico':=0.20%, 'Broomfield County; Colorado':=0.20%,
:
[many items omitted for brevity]
:
```

Note that for 500 records in which each is for a distinct county or county equivalent appears once, we expect each to appear as 0.2%, as is the case and consistent with the report 'original sample size': = 500 on the tag. It is also useful for reporting detailed distributions of clinical, population health, and epidemiological and social data based on age and sex, although these factors were not distinguished in the data available for the present study. However, general central tendency and dispersion descriptors of the distributions of the data by value are reported on the tag for the data detected as quantitative, as follows.

‘Population Health': = '74.32 + / − (17.70P95, 9.03SD, 9.63SDs, 0.40SE, 0.79CI, 0.96Skew, 4.74Kurtosis).

Equity: = '59.12 + / − (30.65P95, 15.64SD, 15.88SDs, 0.70SE, 1.37CI, 0.12Skew, − 2.47Kurtosis)'

Education: = '55.77 + / − (25.29P95, 12.90SD, 13.16SDs, 0.58SE, 1.13CI, 0.18Skew, − 2.13Kurtosis)'

Economy: = '66.54 + / − (24.14P95, 12.32SD, 12.69SDs, 0.55SE, 1.08CI, 0.32Skew, − 1.15Kurtosis)'

Housing: = '56.96 + / − (29.15P95, 14.87SD, 15.10SDs, 0.67SE, 1.30CI, 0.13Skew, − 2.42Kurtosis)'

‘Food & Nutrition': = '62.85 + / − (22.71P95, 11.59SD, 11.94SDs, 0.52SE, 1.02CI, 0.32Skew, − 1.13Kurtosis)'

Environment: = '59.88 + / − (23.41P95, 11.94SD, 12.25SDs, 0.53SE, 1.05CI, 0.26Skew, − 1.56Kurtosis)'

‘Public Safety': = '69.40 + / − (19.38P95, 9.89SD, 10.37SDs, 0.44SE, 0.87CI, 0.64Skew, 1.51Kurtosis)'

‘Community Vitality': = '63.47 + / − (20.71P95, 10.57SD, 10.95SDs, 0.47SE, 0.93CI, 0.42Skew, − 0.35Kurtosis)'

The principal initial findings of socioeconomic interest are also discussed here first to provide an initial overview, and are shown in Tables 1 and 2. However, these initial studies (and Table 3 later below) make relatively simple use of data mining since, for comparison with pairwise Pearson's' correlation, Population Health is explored for trend with other factors just one at a time. Table 1 reports on the 2018 data or overall ranking for the top 500 ranks counties etc. (corresponding to class ‘Overall’), but now joined with the rankings for the top counties etc. ranked in the top 100 of each of the urban high performing, urban up-and-coming, rural high performing, and rural up-and-coming classes.

A more detailed explanation of the calculations here and in Table 1–3 below are given below in this Section, but the main measures reported correspond to standard notions in classical correlation and evidence based medicine, and the general meaning should be intuitive. Except in one study (last row) to make comparison with Table 2, this 2018 data used in Table 1 did not contain joined hospital stay data. There are 800 records in total in which some counties or county equivalents occur twice but are distinguished by e.g. urban up-and-coming and the score within that class for use in data mining. It has the important advantage of including data for a variety of records in which the score called Population Health is often significantly less than 50, rather than a floor of around 50. It covers the range 35–100 such that Population Health ≥50, arbitrarily chosen, was originally reasonably used for predictions in Table 1, but later an optimal threshold of 87 was used. The first value in the correlation Column 2 is for the five combined classes that collectively have some counties or county equivalents counted twice, while the following value in brackets comes from the class ‘Overall’ data in which they are counted only once, and so represent the correlations of more general and traditional socioeconomic (as opposed to methodological) interest. Table 2 is perhaps less confusing as counties did not occur twice in the 2019 data as used, and the numbers in brackets are simply the 2018 numbers on the above “counties only once” basis for comparison. The results are not greatly different especially for results of particular interest such as the negative correlation between Population Health and Equity, meaning that the basic story does not change, but note that the first of the values in the pairs of correlations in Table 1 follow the likelihood ratio, sensitivity, and specificity better. This is consistent with the fact that they are derived from exactly the same data. Nor are they greatly different in combined data where the above threshold or range for Population Heath is taken arbitrarily as 50 but when subject to optimization of which a suitable value was found to be 87. Of course, from a classical population analysis point of view, by representing “top counties” by the

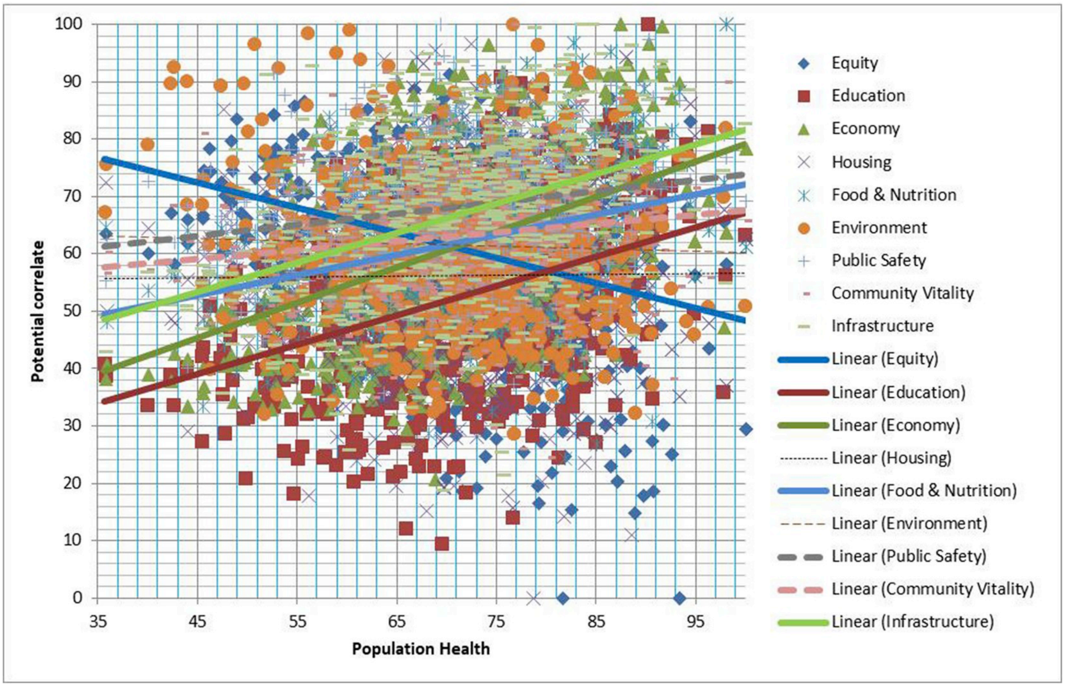


Fig. 2. Scatter plot with regressions for Population Health against its potential determinants in the total combined data.

criteria of that web site [3], the data behind all these Tables 1–3 and the data mining below still represents a rather biased sample.

As well as Q-UEL tags containing more standard analytical information such as the SS method tag above, comma separated value (CSV) files can be generated as input to, for example, a Microsoft spreadsheet. A scatter plot with best fit linear regression lines is shown in Fig. 2 for the original combined 900 records mined in Table 1, which includes some counties or county equivalents twice. This is because (a) scatter plots and graphs encapsulate important features of the data in an introductory way prior to extensive analysis, (b) they are appropriate for comparisons and cross-validations with the general data mining, and (c) because in this case as noted above, the combined classes include a lot of data for which the Population Health is less than 50, covering the range 35–100. As Fig. 2 shows, most factors show a significant positive trend as correlation and regression, three have insignificant trend, but only one, Equity, shows a significant negative trend.

Table 2 discussed next below also gives the Pearson's correlations both for Fig. 2 and for the more limited data in which each county appears only once. While many factors become less significant for the smaller data, the striking negative correlation for Equity is essentially

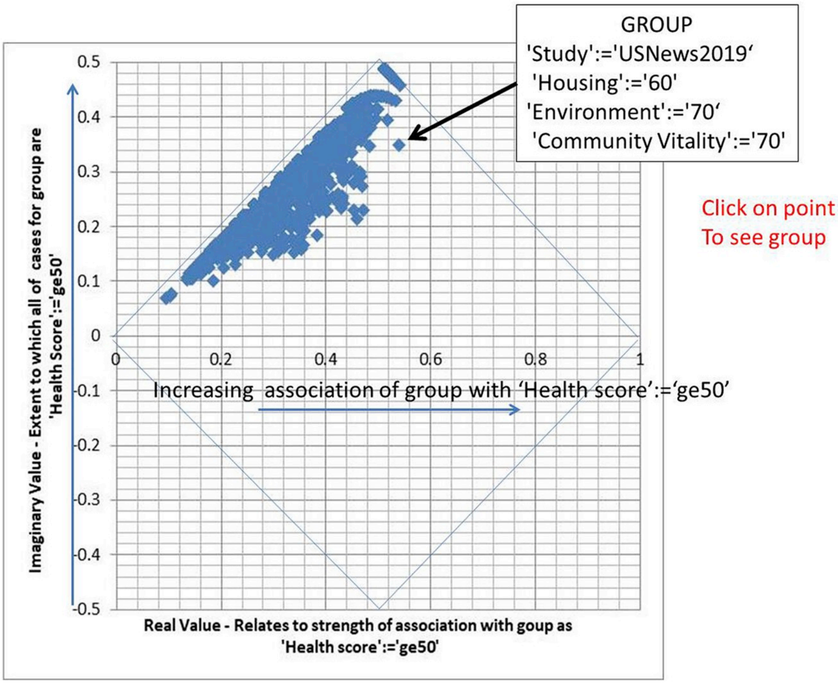


Fig. 3. Graphic showing ***h***-complex probabilities (probability amplitudes, probability duals) of factors associating with Population Health from the mining of the combined data associating factors plotted as a function of their real and imaginary parts.

unchanged. Table 2 looks at the recent 2019 data for the top 500 overall joined with hospital admissions and length of stay data recently accessible to us. In contrast to the 2018 data that was extracted by our methods from the website pages, the data used in Table 2 was downloaded as a file already prepared on the site [3] and did not combine any data from other classes as above. The data mining results by DiracSmash (including likelihood ratio, sensitivity and specificity) shown in Table 2 therefore do not include duplicate counties. An average score for Population Health was 72.5. A prediction of Public Health ≥ 87 was again found to be or close to the optimal threshold or range in most cases, and this was used for the studies in all rows in Table 2 in order to compare results on a similar basis.

The following notions of significance were used. When looking at Pearson correlation values in Column 2 of these Tables, by the usual widely held convention, only negative correlations of -0.2 or less and positive correlations of $+0.2$ or more are considered significant. However, this has no fundamental theoretical basis and correlation, like all the following "statistics", is essentially a matter of degree. For Column 4, a likelihood ratio (of which Relative Risk is an example) of greater than 1 indicates that the factors in Column 1 that are to be compared with population health are more likely to represent low or high scores if the population health score above or below respectively a stated threshold (≥ 87 in Table 1). Based on conventions at least in our data mining, odds of this kind are considered as strongly significant if they represent an absolute information content of 0.5 nat (natural information units), i.e. the likelihood ratio is $e^{+0.5} = 1.6487 \ldots$ or more, or $e^{-0.5} = 0.6065 \ldots$ or less. The theory of expected information used [20] causes the likelihood to start to approach 1 as data becomes increasingly sparse. For sensitivity and specificity in columns 5 and 6 respectively, a value around 50% would be a random prediction, but 60%–70% is considered as having some predictive power. Less than 50% would mean worse than random and usually corresponds to a likelihood ratio less than 1. This actually contains information and is the reason why predictions of a Target are automatically switched to those of the complementary state such as Population Health < 87 with Equity ≥ 56. However, to make comparisons less confusing, we repeat with Population Health ≥ 87 with Equity < 56 in such cases.

In examining Tables 1 and 2, as well as Table 3 below, note that the two-way Pearson's correlation, the likelihood ratio, and the sensitivity and specificity of predictions, are not precisely deducible from each other, but they are for the most part mutually consistent. For example, in Table 1 with results ranked by the size of the likelihood ratio, pairwise correlations of Economy, Infrastructure, Education, and Food and Nutrition with Population health are large and positive when likelihood ratio exceeds 2 and sensitivity and specificity are reasonably balanced (see below) and high. Public safety and Community Vitality are marginal cases by likelihood ratio still exceeding 1, and this is reflected in rather weak correlation, specificity and sensitivity. Equity and length of stay in hospital have negative correlation with Population Health, likelihood ratios less than 1, and rather poor sensitivity and specificity in prediction. While the above consistency between correlation, likelihood ratio and sensitivity and specificity tends to be the case for numerical data, there can be exceptions which tend to come with more complex non-monotonic relationships. Note also that a reasonable likelihood ratio such as for Community Vitality ≥ 60 does not necessarily guarantee very high sensitivity and specificity (Table 1). In Table 2, the general story is the same. The rank order by likelihood ratio has changed, but Economy, Infrastructure, Education, and Food and Nutrition remain the top four, while the negative correlations with Equity are still a strong, and indeed strengthened, feature.

The significant negative correlation of Equity with Population Health that was at first surprising, at least to the authors, provided a motivation for further "drill down" with our less traditional techniques in order to explore further. Length of stay, is considered in Section 4.2. In the interesting case of Equity, Equity ≥ 66 and Equity < 66 were in turn entered initially to demonstrate the effect and the following consideration. In comparing the results of that, note that final predictions from the HDN can vary from exact calculations even with corrections if weak associations are purged from memory, so the likelihood ratio 1.982 is not the exact reciprocal of 0.534 (the exact reciprocal is 1.873 in Table 1) for Equity ≥ 66 and Equity < 66 runs done separately.

4.2. A more detailed description of the statistics and scores generated

Recall that good Population Health was fixed arbitrarily as a score of 50 in initial studies on combined data, but an optimized value 87 or more for the data that based only on higher Population Health (Tables 1 and 2), the other factor involved in each case has a range such as Infrastructure ≥ 60 which is optimized by ROC tuning (see below). This turns out to be essentially the same in practice as tuning to optimize the likelihood ratio, and also with respect to predictive odds, which is proportional to that likelihood ratio by a constant for the given data that is the prior odds, e.g. P(Population Health ≥ 50)/(1 - P(Population Health ≥ 50)). In both Tables 1 and 2, Column 2 shows the classical Pearson's correlation R for Population Health (recall, meaning the population health score) with the metadata (column headings) in the first column, e.g. Economy, and ranges such as ≥ 51 are ignored when considering this classical correlation. The ranges relate to conversion the numerical data to categorical for analysis by associative data mining, which is relevant to the remaining columns. Column 3 relates to size of the large HDN as an estimate generated for comparison purposes: see Section 4.3.

Column 4 shows the calculated likelihood ratio, an example of which is Relative Risk. Predictive odds are also calculated but it is the prediction of Population Health ≥ 50 or Population Health ≥ 87 is made as 'yes' or 'no' according to whether the likelihood ratio exceeds 1, although the ROC curve tuning [26] described below in effect explores the range between Predictive odds and likelihood ratio. These odds are computed by the multiplication of Q-UEL tags (corresponding to brakets, in the theory) obtained by a combination of supervised and unsupervised data mining. As shown in column 3, this process can generate very large inference nets because of the number of columns and the cardinality of the data. To compare the size of our odds HDNs with the much smaller size of typical BNs, note that the number of conditional probabilities involved is twice the number of tags (since each tag represents a dual probability and dual odds) in Column 3. The multiplication process and expected odds shown on the tags represents use of the theory of expected information used to get as much information as possible out of the available data. Though counts may be few for joint events with many factors, many probabilities computed from these can be brought together. Recall that, as in a court of law, a great deal of weak or circumstantial information can accumulate to overthrow a decision that would be made without it. If data for a tag is reasonably large, say 20 counts or more for the joint event described by a tag, then the expected odds closely corresponded.

4.3. Perturbation method and the perturbation correction factor χ

Although the above made relatively simple use of data mining, DiracSmash went through the steps of doing comparative calculations based on construction of large inference nets as estimates, as Column 3 in each of the Tables indicates [26]. Normally this is done because the HDN constructed provides an estimate of a complicated odds calculation (with many factors) that cannot be calculated directly. The Hitlist by definition contains only interdependent factors as necessary interdependent factors as conditions [26], and usually represents a query for which there is adequate data (as is confirmed and reported by DiracSmash). Consequently, an exact calculation of the odds is done, based on the Hitlist as query. Next, a large odds HDN is constructed as an estimate. The latter brings in unstructured data mining to bring in many other combinations of factors that include the Target. The correction factor χ (a multiplier) is now considered that will bring the exact calculation and the HDN as an estimate into alignment. Normally, the Wishlist is subsequently used to represent independent, necessary, individually sufficient factors, albeit with accumulative effect, and odds terms that do not satisfy the Wishlist are removed. The correction factor χ is then applied to the HDN that was built to satisfy both Hitlist and Wishlist. The Wishlist is empty in the case of producing Tables 1–3, but the correction factor χ is of interest in determining the predictive reliability of the data, at least in the context of the approximations that the odds HDN makes (recall again that any inference net is primarily used to make an estimate). Typically χ is less than 1 representing division of the HDN by a factor in the range 1–1.2, and not uncommonly 1.5. Although it is of course undesirable that a risk odds reported as 0.3 should really be 0.2, or more seriously that odds for imminent risk based on strong laboratory data is estimated as say 6 should in actually be 4, that does not affect a binary yes/no prediction based on a strong likelihood ratio much more or much less than 1. In much of the present study the factor χ implies a division by a factor which is typically in the range in range 2–3 (e.g. 2.64 for Population Health ≥87 given Food & Nutrition ≥60) since data is more limited overall and dependent on the theory of expected information and associated algorithmic features [26]. In our experience, if a simple traditional clinical calculation of odds is repeated by including many more patient factors, such discrepancies and worse are common. However, for the smaller set of 372 joined population health and hospital stay records, having an extended number of columns (with full sampling so that a million or more tags can be used to build the HDN), can reach a division factor of 10. This is so in regard to the length of stay data joined with the population health socioeconomic data. It is worth noting that the discrepancies are in our experience of the same order of error *per node* as those in the BN which, being based on the DAG assumes independencies of parent nodes. In the case of the HDN these, and the discrepancies represented by χ are automatically corrected as part of the overall algorithm.

4.4. Quality measures obtained and ROC tuning

Sensitivity and specificity in Columns 5 and 6 also require more explanation as follows. The classical correlation and even the likelihood ratio may not reflect the fullest predictive capability and hence true information content of the data in terms of true positives and negatives and false positives and negatives and hence sensitivity, specificity, etc. when the prediction method is tuned (optimized) by the ROC Receiver operating characteristics approach) applied in DiracSmash [26]. Recall that if the likelihood ratio is less than 1, i.e. the prediction is ‘no’ rather than ‘yes’, then DiracSmash automatically switches the prediction to the complement, e.g. in Table 1 predicting Housing ≥59 is switched to predicting Housing < 59, and predicting Equity ≥66 is switched to Equity < 66, and so on. In general DiracSmash allows a number of different simple prediction models to be tested and this is typically based on the Hitlist interdependent and Wishlist independent factors [26]. However, in this case the matter is somewhat simpler and final sensitivity etc. essentially depend on the prediction as ‘yes’ or ‘no’ as above, on the target Population Health ≥50 being seen on each record with the other factor such as Economy ≥51 for a true positive, and conversely Population Health < 50 and Economy < 51 for a true negative. See Ref. [26] for more details. For most studies in the present project, namely the studies used to produce Tables 1–3, the above effectively zero settings for WishListWeight and DirectionalInfoWeight (See Methods Section 3.3), and a single condition on the Hitlist, meant that in many cases the ROC tuning largely served to reject a Quality threshold value of zero and accept the first non-zero threshold value tested, which was 0.05 (reported as 5%), the same Quality then continuing as a plateau to threshold 1. Note that this does not necessarily imply an equal balance of sensitivity and specificity. In most cases reasonable balance is obtained so that Quality is only slightly less than accuracy which follows the lowest measure. In a case of a very poorly balanced sensitivity and specificity with respect to range and ROC optimization, as obtained for predicting Population Health ≥75 given Economy ≥75, True positives etc. are TP = 85 FP = 40 TN = 197 FN = 182, Accuracy = 56%, Sensitivity = 32%, Specificity = 83%, and Quality = 30%. Generally accuracy lay in the range from sensitivity to specificity inclusive and typically near the mean, and Quality is slightly less than the lowest of sensitivity and specificity. Although accuracy is a popular choice for reporting, it is not a good one. It is well known that it can be misleadingly high in certain circumstances sensitivity and specificity are not balanced, depending on the balance between TP and TN [26].

4.5. Studies on predicting hospital length of stay

As yet, while the data joined with admissions and length of hospital stay gives an interesting negative correlation with Population Health, in this small current data it is as sufficiently weak so as to be considered insignificant by standard criteria. This is supported by rather poor predictive capability to predict Population Health from length of hospital stay and *vice versa*, as shown by sensitivity and specificity. With correlation in the insignificant range, likelihood ratios both close to 1, and sensitivity and specificity close to random (50%). Hence, most of the detailed analyses and results are omitted for brevity here, but will be subject to further study elsewhere because they could be due to competing factors. For example, counties with good population health and economy might allow patients to stay in longer, competing with the possible tendency that longer stays are likely to be reduced by good population health in the county.

Nonetheless, worth mentioning here are other stronger determinants of length of stay found in the 372 joined records that come from the original admissions data, namely the number of admissions annually, as shown in Table 3. Specifying all available current year data 2011–2016 data admissions (for a variety of threshold values for each year) as interdependent necessary factors is not significantly more effective than using just one of these years.

Also, the correlations and likelihood ratios show that the length of hospital stay tends to increase with number of admissions, at least for this small sample of 372 small joined records. This is not the case for the analysis of the original un-joined length of stay records that essentially just comprises number of admission from 2011 to 2016 and annual average length of stay from 2011 to 2016. Preliminary studies indicate that this difference between full admissions data and joined data is *not* simply an artifact due to the much smaller amount of joined data (372 records compared with 2173). It is to be recalled that the county population health data available to us is biased in the sense of only comprising high ranked counties. Consequently, it appears that for those counties that rank relatively highly in scores such as economy and population health, there is a significant, though not strong, correlation, likelihood ratio, and predictive power between number of admissions and length of stay in hospital. The reason for this positive correlation is not obvious though it also appears plausible that the larger hospitals have greater flexibility to accommodate longer stays, and possibly patients in poorer health might tend to be transferred to a larger hospital.

4.6. Results considered in h-complex space

As the above demonstrates, many results obtained by the methodology described here can be described in classical probability and statistical terms, but considerable extra insight and capability comes from the fact that probabilities, predictive odds and likelihood ratios are really expressed in an ***h***-complex space. Graphic representations are often helpful. In Fig. 3, each point is the ***h***-complex value of data-mined Q-UEL tag representing a Dirac braket < A|B > for a particular set of several factors. In this case A is Population Health ≥50 while B will typically be some four other factors, Economy, Education etc. Each Q-UEL tag from data mining associations of the kind considered here carries two probabilities (attributes Pfwd and Pbwd) corresponding to a probability dual represents a probability dual of form {P(A|B), P(B|A)} related to an ***h***-complex value as in Eqn. (10). Importantly for the present paper, the tags also carry odds attributes (predictive odds Ofwd and likelihood ratio Obwd), but probabilities are perhaps easier to understand graphically in regard to approach to limiting values because, just as probabilities have limits 0 and 1, their ***h***-complex counterparts are confined to the so-called "iota diamond" shown in Fig. 3. The corners are bound by the corners with values 0,1, **ι** and **ι***. The abscissa (x-axis) corresponds to ½ [P(A|B) + P(B|A)] as the real or "existential" part of Eqn. (10), representing the probability P("some A are B") which by definition is equal to the P("some B are A″), i.e. the symmetrical component of the relationship between A and B. In this case P(B|A) is small, consistent with the next consideration below, so note that the real value is approximately ½P(A|B) and doubling the value on the diagram gives the real probability (albeit subject to estimation by the use of the zeta function to estimate expected information in this study). Consequently the values correspond to a range of P(A|B) from 0.2 to 1.0 with a majority at about 0.8. The ordinate (y-axis) corresponds to ½ [P(A|B) – P(B|A)] as the imaginary or "universal" part of Eqn. (10), the asymmetrical component representing P("All B are A"), i.e. the extent to which most B are A, from which we can deduce P ("All A are B″), the extent to which most A are B, but it is not the same thing. If these points had been reflected through the abscissa to the lower part of the diagram, one would have conclude the at most A are B. In such cases when the points lie well above or below, the data are said to be "polarized".

The fact that P("All B are A") is here very high, and the majority are as high as possible subject to constraint of the real value, means that the large majority of combinations of factors include high health scores. This is somewhat artificial due to the nature of the data which is for top ranking communities only. We would say that by that the data are "artificially polarized". Nonetheless, that is not obvious without at least some preliminary inspection of that data or a description of it. These kinds of diagrams help us recognize that and interpret accordingly, and also inspection of the points in the diagram as in the box to the upper right of Fig. 3 reveals less obvious, more specific, information. Similar diagrams may be obtained for HDNs as estimates of brakets with many more factors. Nonetheless, a diagram is not particularly helpful for every aspect of interest: often, knowledge of the value suffices. Notably, those HDNs designed to estimate a joint probability should theoretically all be represented by points that all lie on the abscissa, so a single number suffices. Recall that this is because the imaginary part should vanish for the imaginary part of an HDN calculating a joint probability from both directions P(A|B) and P(B|A) of conditional probability, or the corresponding odds, whence the probabilities are said to show coherence with respect to Bayes' rule, but it is rare that it is exactly zero, and in this case the small amount of data meant an important role for estimation of effective probabilities using the zeta function. The joint probability ratios for these studies were typically zero to two decimal places and almost all represented less than 4.0% of the real part of the joint probability ratio, which is considered sufficient for the purpose.

4.7. Studies related to Simpson's paradox

Classical Pearson's correlation studies supported well our data mining and prediction from it, including in regard to the controversial negative trend between Equity and Population Health; however, since Person's correlation and related techniques can sometimes show what is considered a reverse trend due to Simpson's paradox, this demands some consideration. See Fig. 4.

Simpson's Paradox applies when the direction of a relationship at the population-level may be reversed within the subgroups comprising that population, and seems most likely to occur when inferences are drawn across different levels of explanation, from populations to subgroups, or subgroups to individuals [50]. In the case of the present data Simpson's Paradox could apply as shown in the schematic diagram in Fig. 4, in which the overall trend of a putative causative factor with Equity remains undesirably negative, but at least success has been achieved in each group. The paradox is often described as applying when a trend appears in several different groups of data but *disappears* or reverses when these groups are combined. In this study, when all scores except Population Health are added or averaged and plotted against Population Health, the regression slope is 0.10 and correlation becomes insignificant. This is consistent with poor predictive power for BILL that assumes at least a roughly additive effect as discussed in Methods Section 3.3. Note that correlation is also weak in "computer experiments" in which scores (except Population Health) are averaged over just a few factors such as economy as potential determinates of population health are removed one at a time or collectively, giving regressions slopes of 0.10–0.16.

The schematic Fig. 3 is reminiscent of Fig. 2, and Equity is fairly complicated, so making it a possible candidate for Simpson's Paradox because like the other scores it comprises several subgroups combined prior to the public availability of the data. In this case these are Racial Disparity in Educational Attainment, Segregation via Theil Index, heath conditions equity via Air Toxics Exposure Disparity Index Score and Premature Death Disparity Index Score, Income Equity via Gini Index Score and Poverty Disparity Index Score, and Disability Employment Gap [3]. Detailed collated subgroup data is not available at the time of writing but data mining can reveal hidden underlying variables that have impact by partitioning data, here the various scores, into ranges of value that each form a categorical class. As Fig. 2 and particularly Tables 1 and 2 indicate, most factors correlate positively with Population Health. However, that the trend is not in general the same regression. The linear regression is significantly considerably offset meaning that the intercepts would have to be adjusted by a considerable amount to bring them into alignment. The possibility is open that Economy, Education, and Food and Nutrition etc. create separate underlying unknown subgroups, or even that, for example, high (or low) scores for these factors more directly represent fairly distinct groups within which Equity does increase with Population Health. For example, areas with good food and nutrition are not necessarily those with good housing or with good infrastructure, but represent in some way "a different kind" of population. The possibility that Simpson's paradox applies is usually tested by a *segmentation* approach to the data, and associative data mining looks at categorical data and quantitative (numerical) data rendered categorical by putting it into ranges, e.g. 60–69, 70–79 and so on, which can be readily varied. Selected areas of

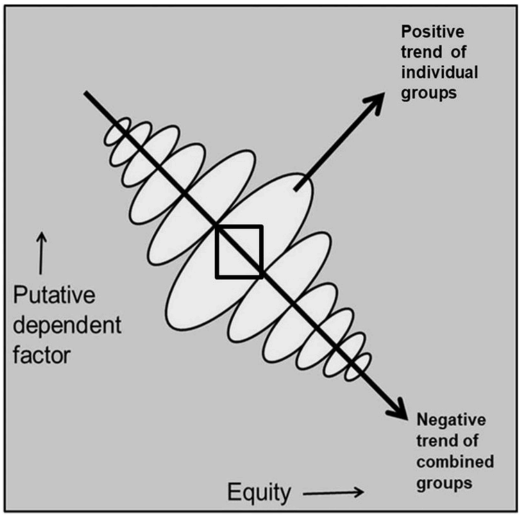

Positive trend of individual groups
Putative dependent factor
Negative trend of combined groups
Equity

Fig. 4. A Purely Schematic Illustration of Simpson's Paradox. The trend behavior within a single group can sometimes by exploring rectangular sections of the scatter plot.

Fig. 2 that can be explored optimized by adjusting binning ranges. It can also be explored by adjusting thresholds of ranges such as 50 in 'Equity': = '**le**50’ (“less than or equal to 50”) in target, Hitlist and Wishlist to other values, by selecting records prior to analysis that contain various ranges, and even by selecting records in which, for example, a putative dependent factor such as economy a related by an equation with positive slope, specifically looking to see if a positive trend within a group is consistent with the counts, probabilities and odds. For brevity only one example tag is given here, pulled out by the predictive odds query PO(Population Health ≥ 50, Equity ≤ 66, rest unsupervised) ≥ 40.

```
< 'Population Health':='ge50' Ofwd:=47.0755 | if:=count:=46 | 'Equity':='le66' 'Community
Vitality':='60-69' 'Public Safety':='70-79' Pbwd:=0.0538 Obwd:=1.4565 >
```

Despite querying for and examining many tags in the above manner, and despite changing and seeking to optimize the score ranges in data mining to segment the data, no strong evidence was found of the overall involvement and impact of Simpson's paradox, at least as described in Fig. 3, when using associative data mining. There are inevitably interesting exceptions, but whether these are usefully interpreted as hints of the paradox is open to debate (See, Discussion and Conclusions Sections 5 and 6). Consequently, the extensive data mining results will be presented elsewhere in a broader discussion of the Equity problem. Essentially, they support the negative Pearson's correlation and the a main finding is that an overall negative correlation is a genuine feature of the current data. In general, predictive odds and likelihood ratios exceeding 1 associated with factors exceeding a critical score, except in the case of Equity, which required 'Equity': = '**le**66’, i.e. Equity ≤66. Higher Population Health with strong predictive odds and likelihood ratios, associate with economy, infrastructure, and education scores in particular. Unsupervised data mining in DiracSmash [26] and DiracMiner [20] shows the above to be associated with each other, and with above-average community vitality, public safety, environment, and food & nutrition (here, found variously in the 50–69 range). In contrast, Equity scores above any range produced no tags with significant positive predictive odds, but they do for Equity considered low (here, equal to or below 66). Purely unsupervised data mining by DiracMiner [20] supports this and shows the above to be associated with each and with above average community vitality, public safety, environment, and food & nutrition (here found variously in the 50–69 range). Generally speaking, associative data mining with an unsupervised component like that above is considered more powerful than *pattern discovery* as follows.

4.8. Pattern discovery tags and “Rule” generation

SMASH optionally also generates pattern tags for purposes of *pattern discovery*. A pattern is simply a combination of factors (in the present study there is no need to consider a particular order), that occurs more than a specified number of times in the data. This number is just 2 in the present study, although the counts of occurrences were often quite high, as shown in the following example tags.

```
<Q-UEL-PATFACTORS-4 'Population Health':='ge50' 'Public Safety':='60-69' 'Community
Vitality':='60-69'   Pfwd:=0.00020553
| if:=count:=49 | 'Equity':='ge66' Q-UEL-PATFACTORS-4>
<Q-UEL-PATFACTORS-4 'Population Health':='ge50' 'Food and Nutrition':='50-59' 'Public
Safety':='60-69'   Pfwd:=0.00020553
| if:=count:=49 | 'Equity':='ge66' Q-UEL-PATFACTORS-4>
<Q-UEL-PATFACTORS-4 'Population Health':='ge50' 'Food and Nutrition':='50-59'
'Environment':='60-69'   Pfwd:=0.00020134 | if:=count:=48 | 'Equity':='ge66' Q-UEL-
PATFACTORS-4>
<Q-UEL-PATFACTORS-4 'Class':='Overall' 'Population Health':='ge50' 'Infrastructure':='70-79'
Pfwd:=0.00019714
| if:=count:=47 | 'Equity':='ge66' Q-UEL-PATFACTORS-4>
<Q-UEL-PATFACTORS-4 'Population Health':='ge50' 'Food and Nutrition':='50-59' 'Community
Vitality':='50-59'   Pfwd:=0.00019295 | if:=count:=46 | 'Equity':='ge66' Q-UEL-PATFACTORS-4>
```

A pattern may be insignificant and coincidental or even occur less than would be expected on a chance basis, and what constitutes patterns can change sensitively with the data. The major objections (e.g. Ref. [52]) on “patterns” as “rules” mostly relate to the latter. However, they sometimes contain many more factors (potential determinants) than those found in association studies, when there are many instances, they can be considered as useful for proposing potential rules worthy of consideration, such as the following. It applies to this study and is consistent with the notion of sub-groups and Simpson's Paradox. “*In records where the equity score is 66 or more, population health can be good providing that two or more of community vitality, food and nutrition, public safety and infrastructure are at a reasonable level possibly reflecting public and administrative care irrespective of economy, education and housing*”. “Rules” like this are, however, also commonly considered as subject to changing with increases in data. An estimated probability Pfwd for such as P(factors | 'Equity': = 'ge66’) does not have the same meaning here as it does in other tags and detailed discussion is beyond present scope, but briefly it is the number of occurrences divided by the total number of all patterns found, and here it is lower than is typically the case. The definitive measures relating to estimated true probabilities and odds are those odds tags in Section 4.7 above. Those showed associations that follow the negative trend of Pearson's correlation for Equity with other factors, and pattern discovery cannot be considered as contradicting that, but it does suggest that high Equity can be positively associated with other scores in certain circumstances and indicates what circumstances are worthy of further investigation.

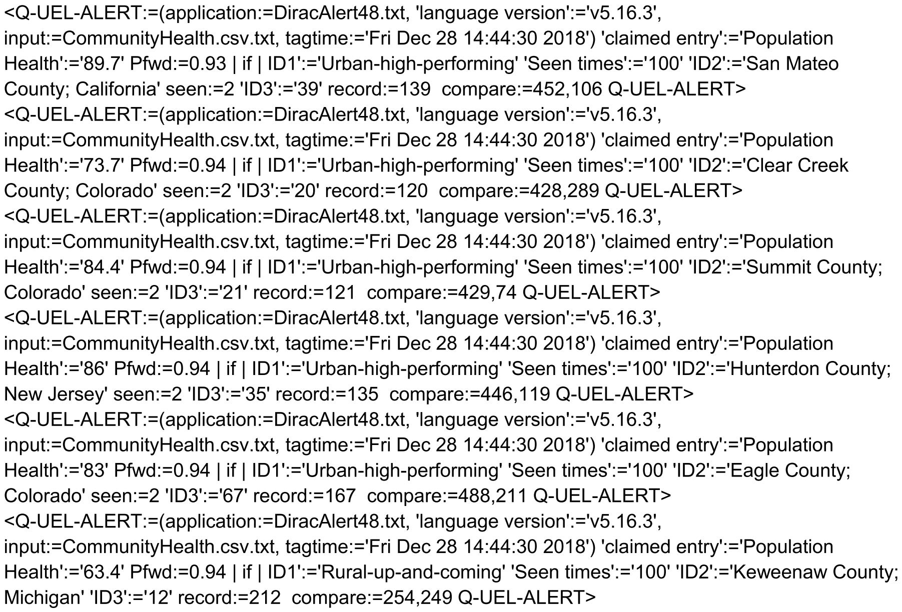
4.9. Anomalous case detection

BILL, ALERT, DiracSmash Normality score [26], and the DiracMiner Patient Analysis mode [20], emphasize anomaly detection. This is concerned with detecting unusual records or unusual features on records, including potentially erroneous or fraudulent records. Anomalous records in the current study may be considered as unusual cases concerning counties, effectively counties that are unusual within the context of the current analysis. They may provide useful information and clues to interpretation because they might be considered outliers. For example, records showing high Equity and high Population Health might provide important clues as to future course of action to improve both. BILL did not prove useful in this particular study (though the fact that it did not find Population Health as not even remotely additive in terms of the other factors is not inconsistent with the idea that Simpson's Paradox might be present). ALERT was initially expected to be inappropriate because it is reasonably expected only to work with very

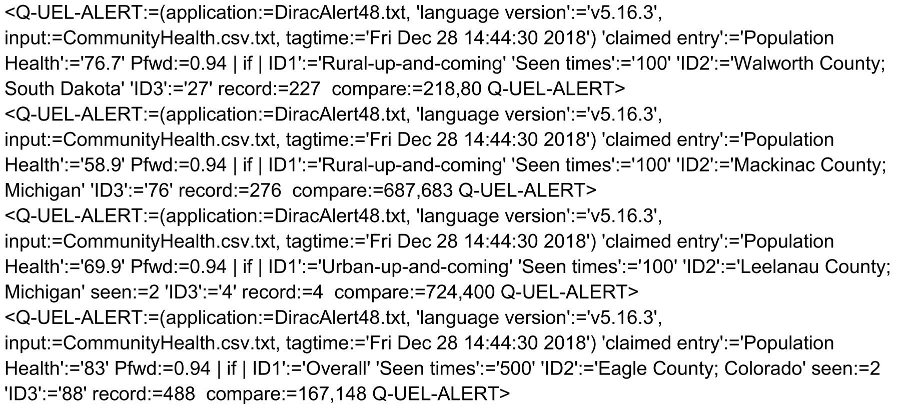
Table 4
Score entries on the records for counties scoring Normality score 1 or less.

County	Population Health	Equity	Education	Economy	Hous-ing	Food & Nutrition	Environ-ment	Public Safety	Community Vitality	Infra-structure
Los Alamos County; New Mexico	**94.4**	83.1	66.5	67.2	86.2	68.7	76.7	72.8	75.5	88.6
Morgan County; Utah	**90.1**	77	63	77.9	55.3	63.5	75	77.7	75.1	63.6
Ozaukee County; Wisconsin	**90**	72.6	77.3	79.8	60.5	65.7	66.1	70.2	71.3	76.4
Lincoln County; South Dakota	**88**	66.2	56	86.8	80.7	50.3	61.6	62.1	79	83.4
Hamilton County; Indiana	**85.1**	69.1	77	91.2	73.2	54	47.8	63.8	83.2	83.2
Delaware County; Ohio	**82.8**	72.9	73.1	88.5	69.6	53.3	53.8	64.1	84.1	79.3
Dukes County; Mass.	63.6	72.5	82.4	68	34.5	71.9	57.2	60.2	91.7	86.5

large numbers of records, but interesting results were obtained. Top-ranking records indicate the most unusual, and the top 10 were as follows. Note first that, in normal use for very large data, Pfwd is very small for unusual records, and the smaller that value, the more unusual the record. Because data was sparse, the following top 10 are also all those records at Pfwd 0.94 or less.



```
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='89.7' Pfwd:=0.93 | if | ID1':='Urban-high-performing' 'Seen times':='100' 'ID2':='San Mateo
County; California' seen:=2 'ID3':='39' record:=139  compare:=452,106 Q-UEL-ALERT>
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='73.7' Pfwd:=0.94 | if | ID1':='Urban-high-performing' 'Seen times':='100' 'ID2':='Clear Creek
County; Colorado' seen:=2 'ID3':='20' record:=120  compare:=428,289 Q-UEL-ALERT>
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='84.4' Pfwd:=0.94 | if | ID1':='Urban-high-performing' 'Seen times':='100' 'ID2':='Summit County;
Colorado' seen:=2 'ID3':='21' record:=121  compare:=429,74 Q-UEL-ALERT>
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='86' Pfwd:=0.94 | if | ID1':='Urban-high-performing' 'Seen times':='100' 'ID2':='Hunterdon County;
New Jersey' seen:=2 'ID3':='35' record:=135  compare:=446,119 Q-UEL-ALERT>
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='83' Pfwd:=0.94 | if | ID1':='Urban-high-performing' 'Seen times':='100' 'ID2':='Eagle County;
Colorado' seen:=2 'ID3':='67' record:=167  compare:=488,211 Q-UEL-ALERT>
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='63.4' Pfwd:=0.94 | if | ID1':='Rural-up-and-coming' 'Seen times':='100' 'ID2':='Keweenaw County;
Michigan' 'ID3':='12' record:=212  compare:=254,249 Q-UEL-ALERT>
```



```
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='76.7' Pfwd:=0.94 | if | ID1':='Rural-up-and-coming' 'Seen times':='100' 'ID2':='Walworth County;
South Dakota' 'ID3':='27' record:=227  compare:=218,80 Q-UEL-ALERT>
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='58.9' Pfwd:=0.94 | if | ID1':='Rural-up-and-coming' 'Seen times':='100' 'ID2':='Mackinac County;
Michigan' 'ID3':='76' record:=276  compare:=687,683 Q-UEL-ALERT>
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='69.9' Pfwd:=0.94 | if | ID1':='Urban-up-and-coming' 'Seen times':='100' 'ID2':='Leelanau County;
Michigan' seen:=2 'ID3':='4' record:=4  compare:=724,400 Q-UEL-ALERT>
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='83' Pfwd:=0.94 | if | ID1':='Overall' 'Seen times':='500' 'ID2':='Eagle County; Colorado' seen:=2
'ID3':='88' record:=488  compare:=167,148 Q-UEL-ALERT>
```

The method requires inspection of the records because the algorithm does not currently distinguish the multiple ways in which records can differ (except by changing setup control for the algorithm used and rerunning). Notably San Mateo county came out as number 8 for median household income out of 3144 county and county equivalents in the United States [53], with an appropriately high economy score of 91.2 in the present data, yet has an extreme gradient of wealth and a very poor Equity of 17.8, putting its entries at 7 and 8 from the bottom for the current Equity data used. For comparison, the records that had a Pfwd of 0.99 or more represented the top 3 that scored as typical, and were as follows. Note that Hawaii County, ranked as most typical of the records of the 'Urban-up-and-coming' class, had a fairly typical Economy of 40–50 and Equity of 50–60, although with Population Health at 77.6 it was scoring well.

```
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='70.5' Pfwd:=0.99 | if | ID1':='Urban-high-performing' 'Seen times':='100' 'ID2':='New Kent County;
Virginia' seen:=2 'ID3':='40' record:=140  compare:=453,438 Q-UEL-ALERT>
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='42.6' Pfwd:=0.99 | if | ID1':='Rural-up-and-coming' 'Seen times':='100' 'ID2':='Swain County;
North Carolina' 'ID3':='78' record:=278  compare:=804,653 Q-UEL-ALERT>
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='78.6' Pfwd:=0.99 | if | ID1':='Urban-up-and-coming' 'Seen times':='100' 'ID2':='Hawaii County;
Hawaii' 'ID3':='62' record:=62  compare:=774,551 Q-UEL-ALERT>
```

Note that Los Alamos County New Mexico with a low Normality Score of 1 and at the top of the list in Table 2 for the DiracSmash Normality score output, discussed below. It was the 75th ranked tag from ALERT, as follows.

```
<Q-UEL-ALERT:=(application:=DiracAlert48.txt, 'language version':='v5.16.3',
input:=CommunityHealth.csv.txt, tagtime:='Fri Dec 28 14:44:30 2018') 'claimed entry':='Population
Health':='94.4' Pfwd:=0.96 | if | ID1':='Urban-high-performing' 'Seen times':='100' 'ID2':='Los Alamos
County; New Mexico' seen:=2 'ID3':='3' record:=103  compare:=403,347 Q-UEL-ALERT>
```

Los Alamos County is notable as unusual by having Population Health ranked in the top 4 at 94.4 and Equity 83.1. Of course, the modules BILL, ALERT and DiracSmash Normality score, as well as the DiracMiner Patient Analysis mode discussed below, are designed to detect different aspects of typical and unusual character, rather than provide redundant information by using similar criteria. With ALERT using the Population Health ≥50 pooling, the Population Health score was in this case deemphasized to mean essentially “above average”. However, ALERT did rank Los Alamos County 8.3% down the list of likely anomalous records by other criteria.

See Table 4. Normality Score is a special mode of DiracSmash because it relies on the results of normal associative data mining. It looks for those records that match the associations of features on the data mining tags, and focuses on those that emerge with least matches and so with lowest probabilities of occurring. That is, the lower the normality score, the more unusual (anomalous) is the record compared with the others in the data mined set. This mode also selects those records that have a normality score equal to or less than one specified by the user. For brevity, class, rank in class, and normality score have been removed from the records found as show in Table 2, but these represent all the records with lowest normality scores, of 1 in this study. DiracMiner PATIENTANALYSIS tags were design to list strongly associated factors for patients and show the probabilities etc. that the patient will have those clinical or demographic values based on all other patients sampled. The probabilities could be low, raising alerts that the patent has unusual clinical values. This readily adapts to entities and entries other than patients. For example while Equity shows an inverse trend with education, plenty of cases such as Montgomery County Public Schools Virginia appear to be very proactive in promoting non-discrimination [51] and like several others do seem to buck the trend having above average education and quite high equity.

```
<Q-UEL-DIRACMINER-PATIENTANALYSIS-3-FACTOR-CHF-SURVEY:=(application:='Perl
version v5.16.3':=DiracMiner158.txt, input:=CommunityHealth.csv, patient#:=3, prior=0,
tagtime(gmt):='Sat Dec  8 13:54:09 2018')
'Education':='60-69'  Pfwd:=Pfzeta:=1.0000
| if:='do all':=assoc:=atomic:=(Kzeta:=4.1846, classical:=24.1086:=2*810000.0000/2*214*157) |
'County':='Montgomery County; Virginia' and 'Equity':='70-79'  Pbwd:=Pbzeta:=0.0193
Q-UEL-DIRACMINER-PATIENTANALYSIS-3-FACTOR-CHF-SURVEY>
```

Virginia appeared often in the pattern and patient analysis tags generated as urban high performing in particular regard to and Education and creditable Population Health, though rather typical in regard to Equity.

4.10. Example use of XTRACTOR with DiracSmash odds tags

In using XTRACTOR to automatically surf and extract information from the Internet, New Hampshire, was prominent in favoring health education [54,55]. New Hampshire Department of Education recognized that health education builds students' knowledge, skills, and positive attitudes about health. Schools there follow NH Minimum Standards for Public School Approval (Ed 306.40) and NH HIV and Health Education Law (RSA 186.11 and RSA 189.10) give the New Hampshire requirements for Health education. New Hampshire promotes a considerable number of best practice points to provide skills-focused instruction that follows a comprehensive, sequential, culturally appropriate K-12 Health education curriculum, and argue for allocating funds and release time to support annual professional development for teachers of health. In the present analysis using DiracSmash with the 2018 data available at the time, for New Hampshire, scores Population Health are high if Equity is high and Education scores are 70 or more, so bucking the negative Equity trend in that specific sense. Note that data is sparse at typical counts of 4 for joint occurrences of all the potential determinants shown in the tag, so that expected Information zeta approach to present more specific cases is important for making what is in effect an “evidence-so-far estimate” of odds.

```
< 'Population Health':='ge50' Ofwd:=5.0610 | if:=count:=4 | 'Equity':='ge66' 'Environment':='40-49'
'Education':='70-79' 'County':='Rockingham County; New Hampshire' Pbwd:=0.005783 Obwd:=0.1566 >
< 'Population Health':='ge50' Ofwd:=5.0610 | if:=count:=4 | 'Equity':='ge66' 'Environment':='40-49'
'Education':='70-79' 'County':='Merrimack County; New Hampshire' Pbwd:=0.00578347Obwd:=0.1566 >
< 'Population Health':='ge50' Ofwd:=5.0610 | if:=count:=4 | 'Equity':='ge66' 'Public Safety':='70-79'
'Education':='70-79' 'County':='Merrimack County; New Hampshire' Pbwd:=0.00578347Obwd:=0.1566 >
< 'Population Health':='ge50' Ofwd:=5.0610 | if:=count:=4 | 'Equity':='ge66' 'Housing':='30-39'
'Education':='70-79' 'County':='Merrimack County; New Hampshire' Pbwd:=0.00578347 Obwd:=0.1566 >
< 'Population Health':='ge50' Ofwd:=5.0610 | if:=count:=4 | 'Equity':='ge66' 'Housing':='30-39'
'Education':='70-79' 'County':='Rockingham County; New Hampshire' Pbwd:=0.00578347 Obwd:=0.1566
>
< 'Population Health':='ge50' Ofwd:=5.0610 | if:=count:=4' | Equity':='ge66' 'Economy':='80-89'
'Education':='70-79' 'County':='Rockingham County; New Hampshire' Pbwd:=0.00578347Obwd:=0.1566 >
< 'Population Health':='ge50' Ofwd:=5.0610 | if:=count:=4 | Equity':='ge66' 'Community Vitality':='60-69'
'Education':='70-79' 'County':='Rockingham County; New Hampshire' Pbwd:=0.00578347 Obwd:=0.1566
>
< 'Population Health':='ge50' Ofwd:=5.0610 | if:=count:=4 | 'Equity':='ge66' 'Public Safety':='70-79'
'Education':='70-79' 'County':='Rockingham County; New Hampshire' Pbwd:=0.00578347Obwd:=0.1566 >
```

[the rest excluded for brevity]

For Virginia Population Health scores are high if Equity is high and Education scores are 30 or more, so also bucking the negative Equity trend in that specific sense, but now more typical with a weaker Education score base of 30. However, counts for joint occurrence of potential determinants is typically only 2, so use of the zeta function is again important for making a evidence-so-far estimate of odds.

```
< 'Population Health':='ge50' Ofwd:=3.0501 | if:=count:=2| 'Equity':='ge66' 'Economy':='50-59'
'Education':='40-49 'County':='Amherst County; Virginia' Pbwd:=0.003486 Obwd:=0.0944 >
< 'Population Health':='ge50' Ofwd:=3.0501 | if:=count:=2| 'Equity':='ge66' 'Public Safety':='80-89'-89
'Education':='40-49 'County':='Rappahannock County; Virginia' Pbwd:=0.0034856 Obwd:=0.0944 >
< 'Population Health':='ge50' Ofwd:=3.0501 | if:=count:=2| 'Equity':='ge66' 'Education':='60-69'
'County':='Montgomery County; Virginia' 'Class':='Overall' Pbwd:=0.0034856 Obwd:=0.0944 >
< 'Population Health':='ge50' Ofwd:=3.0501 | if:=count:=2| 'Equity':='ge66' 'Food and Nutrition':='70-79'
'Education':='30-39'39 'County':='Craig County; Virginia' Pbwd:=0.003486 Obwd:=0.0944 >
< 'Population Health':='ge50' Ofwd:=3.0501 | if:=count:=2| 'Equity':='ge66' 'Education':='60-69'
'County':='Montgomery County; Virginia' 'Class':='Overall' Pbwd:=0.003486 Obwd:=0.0944 >
< 'Population Health':='ge50' Ofwd:=3.0501 | if:=count:=2| 'Equity':='ge66' 'Environment':='50-59'
'Education':='60-69'  'County':='Stafford County; Virginia' Pbwd:=0.003486 Obwd:=0.0944 >
< 'Population Health':='ge50' Ofwd:=5.0610 | if:=count:=4| 'Equity':='ge66' 'Economy':='30-39'
'Education':='70-79 ''County':='Lexington city; Virginia' Pbwd:=0.0057836 Obwd:=0.1566 >
```

[the rest excluded for brevity]

See Table 5 (2018 data, more consistent with date of additional information above). Qualitative and quantitative studies of the relationships between education, socioeconomic factors, geography and so on are, of course, not uncommon (e.g. Refs. [56–62]). The Brenner-Wechsler-Mcmannus (B–W-M) study and scores [56] are of particular interest here by casting light on the comparison between New Hampshire and Virginia. These authors ask "How School Healthy Is Your State?" and provide "a state-by-state comparison of school health practices related to a healthy school environment and health". The score runs from −7 to +9. Top scorers (score 4–9) are Florida, Hawaii, Rhode Island, South Carolina, Delaware, West Virginia, New York, New Jersey. Note the positive correlation of Equity and Education in New Hampshire. Combining their study with our analysis, it appears that Population Heath versus Education correlation is very significantly enhanced for schools with good health education programs, but they only highlight the dilemma with equity. New Hampshire only scores 2 on the B–W-M scale but supports the data mining findings because it does appear to buck the negative equity trend with education, even though negative correlation of equity with population health enigmatically persists. In contrast, Virginia seems a more typical example.

5. Discussion

5.1. Comparing and contrasting our main approach with other methods

In this paper we brought together data mining of both structured (e.g. spreadsheet analysis) and unstructured data mining (i.e. text analytic, natural language processing), and automated construction of probabilistic inference nets that in this study used probabilities derived from the structured approach. However, structured data mining was the tool most used. How might our particular variation on that method be

Table 5
Brenner- wechsler-mcmannus study and initial correlations.

Factor 1	Factor 2	Pearson Correlation for Class "Overall"	Pearson Correlation For B–W-M top scorer States	Pearson Correlation Virginia	Pearson Correlation New Hampshire
Population Health	Equity	−0.324 (negative)	−0.363 (negative)	−0.522 (negative)	−0.498 (negative)
Population Health	Education	+0.179	+0.480	+0.617	+0.469
Equity	Education	−0.357 (negative)	−0.497 (negative)	−0.443 (negative)	+**0.049**

classified? Most broadly described, the main feature used here was categorical data analytics as contrasted with analysis of continuous quantitatively values, although it importantly included analysis of numbers rendered categorical, as ranges. Arguably, major developments in categorical data analysis are relatively recent, and in part due to the rise of Artificial Intelligence AI [63]. Data mining to obtain probabilities with multiple factors [64] is often included in *association* or *associative* data mining [65] (in the present case represented by DiracMiner and DiracSmash), and is often considered as one of such tools of AI, especially when combined with automated inference and prediction. In contrast, the strength of early classical statistics was in the routine statistical analysis of continuous quantitative variables, and this has changed relatively little since the development of "classical" or "frequentist" statistics promoted by Fisher and Neyman in the early twentieth century [66], building on the earlier work of Pearson [67,68] and others.

The well-known chi-squared test [67] is prominent in the exceptions to the above, being an early development due to Pearson, and treating categorical data; there is a relationship to associative data mining, and especially to our approach. Finer classification of the association data mining approach used here is helped by seeing an association constant such as K(A; B) = P(A, B)/P(A) P(B) as, in the limit of indefinitely large data, the ratio O/E where O is the observed and E is the expected frequency of observation (counts)in the chi-square test [67]. We consider mining for probabilities as associative because association in its general sense is about A and B coming together, and note that P(A|B) = P(A)K (A; B) and P(B|A) = P(B)K(A; B). The chi-squared test is based on the chi squared measure $(\text{O-E})^2/\text{E}$ [67]. However, the squaring losses information about whether O is greater than or less than the value expected on a random basis. At least for judging information content as significance of the brakets as probabilistic statements, our methods relate in part to data mining based on exponentials of the zeta function expression ζ(s = 1, O) – ζ(s = 1,E) [69]. This exponential approach O/E in the limit of large data but is also suited to sparse data and belongs to the Theory of Expected Information first early developed for bioinformatics [44]. There is a subclass of our methods based on other values of s that relate ζ to other kinds of surprise measure, and even directly to probabilities [46], but only s = 1 was used here, so strictly speaking it is sufficient to call it a *harmonic series* (1 + 1/2 + 1/3 + …) approach. Unlike those earlier studies, a prime number theoretic approach to data mining [46] was not used and other techniques were employed for managing the combinatorial explosion [20,26]. Also, despite the interest in comparing Pearson correlation, we did not use the earlier technique of reformulating Pearson correlation R values to resemble an association result or "rule" [70,71]. This is because binning in those earlier studies was binary, usually into above or below the average value, which facilitates interconversion with correlation. More elaborate binning that does not presume monotonic trends in the data is best done as in the present study.

Although there is arguably a deep relationship between aspects of our methods and neural nets, it is natural to consider them as very different. When, as in this study, associative data mining is used to build inference networks, it is usually considered as *probabilistic machine learning* [72,73]. Such methods have been described as "top down" in contrast to the "bottom up" approach using (artificial) neural nets [74]. Neural nets currently suffer from the disadvantage that learning is not associated with explicit probabilities but represented by weights obscurely distributed across the network in a way neither reusable for different problems nor easily understood by humans. Fully resolving that problem will ultimately unify top down and bottom up approaches. Currently, neural nets are best at recognition and categorization, e.g. of faces, with each set of weights appropriate to each specific task, but there is the argument that the human brain uses both bottom up and to down approaches, and hence so should advanced AI [74]. *Deep Learning* is at present enjoying considerable popularity where "deep" usually refers to the introduction of more hidden layers of nodes in neural nets [75]. However, in the history of neural nets, extra hidden layers were always an option, and many authors including ourselves consider that "deep learning" also applies to refinements in probabilistic machine learning, especially for data mining of high dimensionality including cardinality, and large inference nets built from it.

5.2. Comparing and contrasting the performance of our approach with others

Especially because our approach is somewhat unusual, it is comparison between final predictive performances of methods that is easiest to quantify. Unfortunately for that purpose, workers in the socioeconomic field have tended to use "classical" statistical methods and in a descriptive rather than predictive way, even if the data is large [76]. More recently some such projects have used data mining (e.g. Refs. [77–79]), but the focus is not in general on new methods and comparing their predictive power. There have been several relevant studies using medical records. One of the authors (BR) reviewed these and argued a reasonable lead to the HDN in predicting adverse events such as congestive heart failure and renal failure [26] (see also ref [14]). The problem still remains that comparison is difficult, because even if one treats a method as a "black box" and focuses on input data and results, data used are significantly different in nature and size in each study. Some predications by other workers were more of the nature of diagnoses than predictions of future events, the former generally being considered easier because clinical results are supposed to provide strong clues. Some authors report only accuracy, and a balanced high score for both sensitivity and specificity is important because a high accuracy can be obtained by allowing very low sensitivity or a very low specificity. A recent review of neural net Deep Learning and predictions of adverse events and future disease states from medical records, indicated accuracy as high as 79% [80]. A preprint by some of the same authors [81] describing their own Deep Learning approach (DeepPatient) for a selection of some of the serious diseases that the studied reported areas under the ROC curve of 85%–91%. This suggests accuracies of similar values with a more balanced sensitivity and specificity of similar value. While fair comparison remains difficult, this seems comparable with values of 91% accuracy that we obtained. These results were for the most successful out of 5 different sets of clinical and demographic factors as predictors of congestive heart failure and similarly 91% out of 3 such sets for renal failure (See Table 2 in Ref. [26]), noting here that "most successful" is not arbitrary but identifies the relevant clinical and demographic factors. Example strong prediction results of the present paper for predicting Population health from Economy at 80% sensitivity and 91% specificity corresponding to 85% accuracy, and Population Health from Infrastructure at correspondingly 72%, 81% and 75%. However, with very different data and prediction, there is no particular significance claimed for this comparison except to note that these kinds of scores indicate useful predictive capability.

5.3. Relative merits of including our kind of approach in routine statistical studies

Techniques such as neural nets including Deep Learning can classify or recognize complex features in milliseconds but training on standard processors is usually measured in hours or days. Even with the availability of the state of-the-art high-performance computing systems, it can require days to train deep neural networks [82–84]. For example, it took six days to train AlexNet, one of the famous systems that established Deep Learning [82], on two GTX 580 3 GB processors [83]. Our own approaches, applied to large clinical or healthcare insurance claims data of many columns and millions of records, can also take minutes or hours according to detail and complexity required, but it is fast for the present kind of socioeconomic data, even including the combinatorial explosion discussed in Section 1.2.3. For studies by

DiracSmash in the present paper, training and overall testing including ROC curve optimization all from start to finish took 10 s for the longest run in this present study on a relatively slow (2.30 GHz) processor, although for very high dimensional medical data, runs are typically around 44 min [26].

5.4. Comments on correlation, likelihood ratio, sensitivity and specificity

The correspondence between Pearson's correlation and the likelihood ratio, sensitivity and specificity (when derived from exactly the same data) not only validates the data mining and inference net algorithms used but also supports the more general idea that these last three numbers should be high on the list of promising candidates for use in routine statistical studies. Hunting for an involvement of Simpson's paradox is not primarily a matter of looking for agreement or discrepancy between correlation and our current methods. Rather, it is a matter of examining data-mined tags (including those that are used together in large numbers for the purpose of inference) and looking for contradictions. Though no strong support for Simpson's paradox was found in the present study, the paradox is a matter of degree and even hints of it or of non-monotonic behavior can provide clues as discussed in Section 6.5. DiracSmash displays many more measures than likelihood, sensitivity and specificity that we could have reported and used [26], and DiracBuilder enables the user to create many more measures as simple examples of inference nets [20]. However, other measures added little to our present and previous [26] study. Predictive odds is equal to likelihood ratio multiplied by prior odds (Eqn. (15)). This makes the likelihood ratio less variable with proportions of factors in the sample (which dictate prior odds), and so more readily transferable to other populations and to the specific case such as a patient [26] or by analogy a county. In any event it is effectively adjustable between predictive odds and likelihood ratio by ROC curve tuning, essentially adjusting a contribution from prior odds. The advantage of having both sensitivity and specificity is that the user may wish to reset control parameters to favor one over the other, in the manner of optimizing the utility of a decision process. For example, the potential benefits of current prostate cancer testing outweigh the stress caused by a positive result and the harms of further testing and treatment [85]. However, because of the traditional use of standard methods and ideally unbiased reporting in the manner of a census, deliberate under or over prediction is less likely to be a feature of socioeconomic study.

6. Conclusions

6.1. Primary conclusions

The combined use of tools and modes of use described in this paper appears capable of adding significant value to the analysis of socioeconomic health data. Because the significant negative correlations between scores for equity and population health, economy etc. were unexpected at least to the authors, confirmation by several techniques and measures, including long-established Pearson's correlation, was particularly important for us. So was some consideration of Simpson's paradox that can arise in Pearson's correlation, i.e. that a trend appears in several different groups of data but reverses when these groups are combined. No strong evidence for extensive appearance and impact of the paradox was found in this study, but as noted above, its presence is a matter of degree, and occasional hints of it can provide important clues (see Section 6.4 below). In general in statistics it is obvious that Simpson's paradox can matter, but ultimately an alternative argument is possible here. If a negative correlation of Equity with other factors applies overall, then socioeconomically it is an unsatisfactory situation whether Simpson's Paradox applies or not. That is, even if trends with Equity are positive within a group, the overall trend of the groups is unsatisfactory. Also, unlike, say, rise of cholesterol with age in a study of patients receiving a new statin drug, an example of Simpson's Paradox in which study age would not be a factor that could be influenced, any underlying trend between subgroups may in the Population Health case might well be susceptible to curative action. The general conclusion would still remain that either the equity score is imperfect or its goal or equal or appropriate opportunity is not yet achieved.

6.2. Ongoing and future development of the techniques

We are conscious that our tools need to be fine-honed to deal with Simpson's Paradox. In part, this is a relatively simple matter of smarter control of segmentation and binned value ranges in quantitative data. However, despite the lack of evidence obtained for the appearance of Simpson's paradox as it is usually defined, the findings did highlight that there is not a single yes/no answer. A single statistical measure of the *extent* to which the paradox applies would be of considerable value, and additional information as to which parts of the statistical space, say as displayed in Fig. 2, are affected, would be extremely useful. This is all currently underway, but so far it has not produced any significantly different conclusions for the socioeconomic health data studied here. Some problems are less readily surmountable at this moment in time. One is of course simply that the data are very small and further and finer-grained data will probably be needed in order to confirm the findings so far. In using the Theory of Expect Information, information and probability typically start to converge closely to classical values based on ratios of counts after approximately 20 observations. This seems consistent with Fisher's recommendation of $\alpha = 1/20$ (the threshold of the P value for rejecting the null hypothesis) that is popularly said to be based on the intuitive sense of 1 in 20 contrary observations as marginal evidence for concern [66]. Be that as it may, while a few counts of associating events were in the 10–20 range, most were less than 10. Significantly larger counts are not likely to be delivered for the kind of US County data in the foreseeable future, because counties and equivalents are not likely to change much. This boosts the argument for the use of a Theory of Expected Information for sparse data for socioeconomic studies of this kind. There is a more fundamental problem in regard to what one means by "explanation" of relationships found. The kind of tentative "rules" that can be generated about what governs the relationships can be quite complex and it is typical that the more coverage any rule or rule set provides, the more complicated it becomes. A very complex statement is not likely to be a pleasing answer, any more than a long and complicated mathematical proof is pleasing to mathematicians. It is usually possible to give a reasonable synopsis of how a criminal trial reached a conclusion even in quite complex cases, but data mining and/or HDN construction can involve hundreds of thousands or millions of tags as pieces of evidence. Nonetheless, a system that gives some kind of explanation is always preferable. We are building a system in which the knowledge extracted could engage not only physicians but the community and its administrators by probabilistic semantics that is intrinsically easy to understand, test and create, by having been built on a human language model, as well as a common data model [86–88]. Already in clinical applications the basic knowledge elements can be ranked by strength and, when displayed in simplified Q-UEL, are easily directly readable by eye, e.g. < uveitis Pfwd: = 0.3 | **if** | 'ankylosing spondylitis' Pbwd: = 0.2 > , and < uveitis | **'does not necessarily indicate'** | 'Intervertebral disc infection' > .

6.3. Counterexamples and hints of Simpson's paradox can provide clues to socioeconomic influences

A counterexample of the overall negative trend of Equity with Population Health etc. can be suggestive of the nature of the socioeconomic problem and how it might be fixed. For example, a promising "rule" extracted by pattern discovery was that "*In records where the equity score is 66 or more, population health can be good providing that two or more of community vitality, food and nutrition, public safety and* *infrastructure are at a reasonable level possibly reflecting public and administrative care irrespective of economy, education and housing*". As found in Results Sections 4.9 and 4.10, education also sometimes emerges as a potentially important factor. For brevity a more detailed analysis of other will be done elsewhere, and some obviously suggest an appearance of Simpson's paradox in certain parts of the data. For example, New Hampshire had a positive correlation between Equity and Education, and this provides an example of reversal of trend within a subgroup. It is the relatively rare case of Education and Equity not correlating negatively and in data mining. However, nothing suggests to us that this is better seen in terms of Simpson's paradox as imposed to, say, the appearance of informative outliers in the data, and either way, it provides a possible clue worthy of further investigation. Although Education was not always the most prominent among factors which associate with high Equity and Population Health, it is plausible that this is somehow a stronger factor in the problem detected, e.g. current planning for equity might have worse-than-little effect on population health score because it needs increasing equity in education. Also, introducing a strong health guidance syllabus for all may raise the influence of education to that of a major determining factor for population health.

6.4. Some comments on equity and education

The above suggests a potential role for education in enhancing equity. Of all contributing factors that might be addressed to enhance equity, education seems the easiest to implement, involve by far the least cost, and seems inherently unlikely to risk any negative consequences. The Center for Disease Control Healthy Schools Program may provide further useful data that will be explored [55]. It supports programs administered by states that focus on improving the well-being of youth through healthy eating, physical education and physical activity, reducing risk factors associated with childhood obesity, and managing chronic health conditions in schools. It may not be possible to separate out equity, education and health, and they may need to be treated fundamentally as "a bundle". European studies tend to see equity, education and health as an integrated holistic whole, and highlight *transitions between schools*, as well as bullying and sense of some kind of inequality and deficiency, as challenging to health (e.g. Refs. [57–59]). US researchers are also "on the case" (e.g. Ref. [60]), and many look both locally and internationally for an understanding (e.g. Refs. [61,62]). Although further study is required, the experiment of enhancing equity in education and importantly equity in health education is indicated, would seem a worthy aim in itself, and as noted above, would be relatively cost-effective and highly unlikely to have any negative consequences.

6.5. Concluding Remarks

It remains that the authors are cautious about making such declarations outside our area of main expertise, but for three main interesting reasons. First and most basically, not only was the approach somewhat unusual (though supported by classical correlation studies), but also the data comprised scores developed by the data preparers, not the present authors. Hence the impact of this requires further investigation by specialists including those skilled in use of socioeconomic metrics. Second, while our focus has indeed been methodological, to help extend the AI toolkit, the state of the art of AI is still such that predictions by any such methods must be supported by validation [89]. The Alan Turing Institute in London [90] makes a particular study of what is required for safe and ethical AI. Incidentally but remarkably, it is highly appropriate for the present paper that the first item on the list on their site as examined earlier in 2019 was *fairness*, i.e. measuring and mitigating inappropriate bias against subgroups. It was followed by transparency, i.e. improving our understanding of how algorithmic systems operate with a view to their applications in the real world, and by matters of privacy, robustness, resilience, and control to avoid unintended behaviors. Although the list and its details have changed from time to time, fairness as lack of discrimination currently remains a first priority of the Institute at the time of writing [90]. In other words, matters related to equity are considered as major issues for the development of AI. Third and finally, note that our findings for population health and economics etc. with respect to equity, from the US data used, are not consistent with findings on economics and equity more globally. At the global level, on average, increases in the level of income inequality lead to lower transitional GDP per capita growth, and increases in the level of income inequality have a negative long-run effect on the level of GDP per capita [91], supported by a number of studies [92–95]. None of this detracts from the desirability of improving equity, not simply from an ethical perspective but also from an overall economic one, and again education emerges as of interest. Economists appear to accept that "equity-enhancing policies, particularly such investment in human capital as education, can, in the long run, boost economic growth, which, in turn, has been shown to alleviate poverty" [95].

References

[1] B. Robson, O.K. Baek, The Engines of Hippocrates. From the Dawn of Medicine to Medical and Pharmaceutical Informatics, Wiley, 2009.

[2] S.B. Johnson, The Ghost Map: the Story of London's Most Terrifying Epidemic – and How it Changed Science, Cities and the Modern World, Riverhead, (2006).

[3] https://www.usnews.com/news/healthiest-communities last access 4/20/2019.

[4] 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, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf, (2010) last accessed 12/16/2018.

[5] P.A.M. Dirac, A new notation for quantum mechanics, Math. Proc. Camb. Philos. Soc. 35 (3) (1939) 416–418.

[6] B. Robson, The new physician as unwitting quantum mechanic: is adapting Dirac's inference system best practice for personalized medicine, genomics and proteomics? J. Proteome Res. 6 (No. 8) (2007) 3114–3126.

[7] I.M. Mullins, I.M., M.S. Siadaty, J. Lyman, K. Scully, G.T. Garrett, G. Miller, R. Muller, B. Robson, C. Apte, S. Weiss, I. Rigoutsos, D. Platt, S. Cohen, Data mining and clinical data repositories: insights from a 667,000 patient data set, Comput. Biol. Med. 36 (12) (2006) 1351.

[8] B. Robson, Towards Intelligent Internet-Roaming Agents for Mining and Inference from Medical Data, Future of Health Technology Congress, Technology and Informatics vol 149, IOS Press, 2009, pp. 157–177.

[9] B. Robson, Links between Quantum Physics and Thought (A. I. Applications in Medicine), Future of Health Technology Congress, Technology and Informatics vol 149, IOS Press, 2009, pp. 236–248.

[10] B. Robson, Towards Automated Reasoning for Drug Discovery and Pharmaceutical Business Intelligence, Pharmaceutical Technology and Drug Research, 2012.

[11] B. Robson, Towards new tools for pharmacoepidemiology, Adv. Pharmacoepidemiol. Drug Saf. 1 (6) (2013), https://doi.org/10.4172/2167-1052.100012.

[12] B. Robson, Hyperbolic Dirac nets for medical decision support. Theory, methods, and comparison with Bayes nets, Comput. Biol. Med. 51 (2014) 183–197.

[13] S. Deckelman, B. Robson, B. Split-Complex Numbers and Dirac Bra-Kets vol. 14, Communications in Information and Systems (CIS), 2015, pp. 135–149 3.

[14] B. Robson, Bidirectional general graphs for inference. Principles and implications for medicine, Comput. Biol. Med. 108 (2019) 382–399.

[15] B. Robson, U.G.J. Balis, T.P. Caruso, Considerations , for a universal exchange language for healthcare, Proceedings of 2011 IEEE 13th International Conference on e-Health Networking, Applications and Services (Healthcom 2011), vols. 173–176, IEEE, Columbus, MO, 2011.

[16] B. Robson, U.G.J. Balis, T.P. Caruso, Suggestions for a web based universal exchange and inference language for medicine, Comput. Biol. Med. 1 (12) (2013) 2297–2310 43.

[17] B. Robson, T.P. Caruso, A universal exchange language for healthcare MedInfo '13, in: C.U. Lehmann, E. Ammenwerth, C. Nohr (Eds.), Proceedings of the 14th World Congress on Medical and Health Informatics, Copenhagen, Denmark, IOS Press, Washington, DC, USA, 2013.

[18] B. Robson, T.P. Caruso, U.G.J. Balis, Suggestions for a web based universal exchange and inference language for medicine. Continuity of patient care with PCAST disaggregation, Comput. Biol. Med. 56 (2014) 51–66.

[19] B. Robson, POPPER, a simple programming language for probabilistic semantic inference in medicine, Comput. Biol. Med. 56 (2014) 107–123.

[20] B. Robson, S. Boray, 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. 66 (2015) 82–102.

[21] B. Robson, S. Boray, 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), Proceedings 2015 IEEE International Conference, vols. 1397–1404, IEEE, 2015.

[22] B. Robson, S. 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. 73 (2015) 71–93.

[23] B. Robson, 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, Comput. Biol. Med. 79 (2016) 299–323.

[24] B. Robson, S. Boray, Studies of the role of a smart web for precision medicine supported by biobanking, personalized medicine, FTG, Pers. Med. 13 (2016) 4.

[25] B. Robson, S. Boray, Methods and Systems of a Hyperbolic-Dirac-Net-Based Bioingine Platform and Ensemble of Applications, (2017) US20170185729A1.

[26] B. Robson, S. Boray, 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. 95 (2018) 147–166.

[27] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, San Mateo, 1988.

[28] E.R. Harold, W.S. Means, XML in a Nutshell, O'Reilly, 2004.

[29] Position statement from the workshop on RDF as a universal healthcare exchange language held at the 2013 semantic technology and business conference, san Francisco, Yosemite Manifesto on RDF as a universal healthcare exchange language, http://yosemitemanifesto.org/, (2012).

[30] https://en.wikipedia.org/wiki/Bayesian_network last accessed 7/6/2019.

[31] P.A.M. Dirac, The Principles of QM, fourth ed., Oxford University Press, 1958.

[32] J. Bircher, E.G. Hahn, Applying a complex adaptive system's understanding of health to primary care, F1000 Res. 5 (2016) 1672 www.ncbi.nlm.nih.gov/mc/articles/PMC5043445/.

[33] J.E. Stiglitz, The rigged equality, Sci. Am. 319 (5) (2018) 50–55.

[34] R.M. Sapolsky, The health-wealth gap, Sci. Am. 319 (5) (2018) 56–61.

[35] V. Eubanks, Automating bias (how algorithms designed to alleviate poverty can perpetuate it instead), 319 (5) (2018) 62–65.

[36] J.K. Boyce, The environmental cost of inequality, 319 (5) (2018) 66–71.

[37] http://sgba-resource.ca/en/concepts/equity/distinguish-between-equity-and-equality/.

[38] https://www.diffen.com/difference/Equality-vs-Equity.

[39] http://www.publichealthnotes.com/equity-vs-equality/.

[40] Organization for Economic Collaboration And Development, Educational Opportunity for All, (2017).

[41] The Legatum institute, https://www.prosperity.com/rankings last accessed 6/4/2019.

[42] Emile Durkheim, The Division of Labour in Society, Trans. W. D. Halls, intro. Lewis A. Coser Free Press, New York, 1997 39, 60, 108.

[43] J.J. Gerber, L.M. Macionis, Sociology, 7th Canadian ed., Pearson Canada, Toronto, 2010, p. 97.

[44] B. Robson, Analysis of the code relating sequence to conformation in globular proteins: theory and application of expected information, Biochem. J. 141 (1974) 853–867 1974.

[45] The GOR Method, Wikipedia https://en.wikipedia.org/wiki/GOR_method last accessed 6/4/2019.

[46] B. Robson, Clinical and pharmacogenomic data mining: 3. Zeta theory as a general tactic for clinical bioinformatics, J. Proteome Res. 4 (2) (2005) 445–455.

[47] K. Popper, The Logic of Scientific Discovery, Springer, 1934 as Logik der Forschung; English translation 1959.

[48] https://en.wikipedia.org/wiki/Semantic_triple last accessed 6/4/2019.

[49] W.J. Krzanowski, D.J. Hand, ROC Curves for Continuous Data, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 2009.

[50] R.A. Kievit, W.E. Frankenhuis, L.J. Waldorp, D. Borsboom, Simpson's paradox in psychological science: a practical guide, Front. Psychol. 4 (2013) 513.

[51] Montgomery county public schools, https://www.mcps.org last accessed 6/25/2019.

[52] https://www.healthcareitnews.com/blog/solving-claim-overpayment-conundrum last accessed 6/25/2019.

[53] https://en.wikipedia.org/wiki/List_of_highest-income_counties_in_the_United_States#American_Community_Survey last accessed 6/4/2019.

[54] https://www.education.nh.gov/instruction/school_health/health_coord_education.htm last accessed 6/4/2019.

[55] https://www.cdc.gov/healthyschools/stateprograms.htm last accessed 6/20/2019.

[56] N.D. Brenner, H. Wechsler, T. McManus, How school healthy is Your state? A state-by-state comparison of school health practices related to a healthy school environment and health education, J. Sch. Health 83 (10) (2013) 743–749.

[57] Equity, education and health: learning from practice, in: K. Buijs, Dadaczynski A. Schultz, T. Vilaça (Eds.), Case Studies of Practice Presented during the 4th European Conference on Health Promoting Schools Odense, Denmark, 2013 7– 9 October http://www.schools-for-health.eu/uploads/files/Innovative%20Practice%20Book.pdf.

[58] Dept. Heath and UCL institute of health equity, improving school transitions for health equity, www.instituteofhealthequity.org/resources-reports/improving-school-transitions-for-health-equity/improving-school-transitions-for-health-equity.pdf last accessed 6/4/2019.

[59] S. Anderson, et al., Appropriate settings for health promotion (e.g. Schools, the workplace)in The Public Health Handbook, https://www.healthknowledge.org.uk/public-health-textbook/disease-causation-diagnostic/2h-principles-health-promotion/appropriate-settings-hp last accessed 6/4/2019.

[60] M.J. Blank, Building sustainable health and education partnerships: stories from local communities, J. Sch. Health 85 (11) (2015) 810–816.

[61] L.V. Adams, C.M. Wagner, C.T. Nutt, A. Binagwaho, The future of global health education: training for equity in global health, BMC Med. Educ. 296 (2016), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5117699/ last accessed 6/7/2019.

[62] S. Kumar, I. Chong, Correlation analysis to identify the effective data in machine learning: prediction of depressive disorder and emotion states, Int. J. Environ. Res. Public Health 15 (2018) 2907.

[63] M. Tegmark, Life 3.0: Being Human in the Age of Artificial Intelligence, Penguin Books, 2017.

[64] J. Han, M. Kamber, J. Pei, Data Mining. Concepts and Techniques, Elsevier, 2012.

[65] C. Zhang, S. Zhang, Association Rule Mining, Springer-Verlag, 2002.

[66] E.L.L. Lehmann, Fisher, Nyman, and the Creation of Classical Statistics, Springer, 2011.

[67] R.L. Plackett, Karl Pearson and the chi-squared test, Int. Stat. Rev. 51 (1983) 59–72.

[68] K. Pearson, On Further Methods of Determining Correlation, Dulau and Co., London, 1907.

[69] B. Robson, Clinical and Pharmacogenomic Data Mining. 1. The generalized theory of expected information and application to the development of tools, J. Proteome Res. 283–301 (2003) 2.

[70] B. Robson, R. Mushlin, Clinical and pharmacogenomic data mining. 2. A simple method for the combination of information from associations and multivariances to facilitate analysis, decision and design in clinical research and practice, J. Proteome Res. 3 (4) (2004) 697–711.

[71] B. Robson, Clinical and pharmacogenomic data mining: 4. The FANO program and command set as an example of tools for biomedical discovery and evidence based medicine, J. Proteome Res. 7 (9) (2008) 3922–3947.

[72] K.P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012.

[73] Z. Gharhamani, Probabilistic machine learning and artificial intelligence, Nature 521 (2015) 452–459.

[74] A. Copnik, Making AI more human, Sci. Am. 316 (6) (2017) 54–59.

[75] Y. Bengio Goodfellow, A. Courville, Deep Learning (Adaptive Computation and Machine Learning Series), The MIT Press, 2016.

[76] A. García-Altés, D. Ruiz-Muñoz1, C. Colls, M. Mias, N.M. Bassols, Socioeconomic inequalities in health and the use of healthcare services in Catalonia: analysis of the individual data of 7.5 million residents, J. Epidemiol. Community Health 72 (10) (2019).

[77] L. Hosseini, V. Hossein Vatanpour, M. Mohammadzadeh, M. Abdolahi, Rita Motidostkomleh, Data mining approach for exploring socioeconomic patterns in cancer, Bali Med. J. 7 (1) (2018) 97–103.

[78] S. Vinnakota, N.S.N. Lam, Socioeconomic inequality of cancer mortality in the United States: a spatial data mining approach, Int. J. Health Geogr. 5 (9) (2006).

[79] J. Gonçalves, B.M. Faria, L.P. Reis, V. Carvalho, A. Rocha, Data Mining and Electronic Devices Applied to Quality of Life Related to Health Data, IEEE 10th Iberian Conference on Information Systems and Technologies, 2015.

[80] V. Osmani, L. Li, M. Danieletto, B. Glicksberg, J. Dudley, O. Mayora, Cornell university https://arxiv.org/abs/1804.01758, (2018) last accessed 6/23/2019.

[81] R. Miotto, L. Li, J.T. Dudley, Deep Patient: an Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records, (2016) https://www.semanticscholar.org/paper/Deep-Patient%3A-An-Unsupervised-Representation-to-the-Miotto-Li/18c39ba04333d31c6cb10faf79d1f18692c38d0f/ last accessed 6/23/2019.

[82] J. Yasaswi, S. Purini, C.V. Jawahar, Predicting the training time of deep neural networks, NIPS Proceedings of the 25th International Conference on Neural Information Processing Systems, 1 2012, pp. 1097–1105.

[83] AlexNet, https://en.wikipedia.org/wiki/AlexNet last accessed 6/17/2019.

[84] https://researchweb.iiit.ac.in/~jitendra.katta/papers/predicting_training_time_final.pdf last accessed 6/25/2019.

[85] S. Begley, Should you get screened for prostate cancer? New guidance updates a 2012 recommendation, scientific American, STAT: the body, april, https://www.scientificamerican.com/article/should-you-get-screened-for-prostate-cancer/, (2017) last accessed 6/17/2019.

[86] Common data model, https://pcornet.org/pcornet-common-data-model/ last accessed 6/7/2519.

[87] Informatics for integrating biology and the bedside, https://www.i2b2.org/ last accessed 6/25/2019.

[88] http://chime.ucsf.edu/observational-medical-outcomes-partnership-omop last accessed 6/25//2019.

[89] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Pearson, 2009.

[90] The Alan Turing Research Institute, British Library, London, https://www.turing.ac.uk/research/interest-groups/fairness-transparency-privacy.

[91] M. Brueckner, D. Lederman, Effects of Income Inequality on Aggregate Output vol 7317, (2015) World Bank Policy Discussion Paper.

[92] O. Galor, Inequality, Human Capital Formation, and the Process of Development, Brown University working papers, 2011, pp. 2011–2017.

[93] J.D. Ostry, A. Berg, G.D. Tsangarides, Redistribution, Inequality, and Growth, IMF Staff Discussion Note No. SDN/14/02, (2014).

[94] R. Perotti, Growth, income distribution, and democracy: what the data say? J. Econ. Growth 1 (2) (1996) 149–187.

[95] IMF Fiscal Affairs Department, Should equity Be a goal of economic policy? https://www.imf.org/external/pubs/ft/issues/issues16/, (1999) last accessed 6/20/2019.