Evidence-Based Gene Prioritization

Matevž Kovačič*†‡

Abstract

We propose (1) an evidence-based gene prioritization algorithm based on a Bayesian diagnostic model. Our results suggest that (2) proband phenotype and the HPO ontology alone are sufficient to achieve superior diagnostic performance in a representative group of clinical cases. Furthermore, the algorithm provides (3) the explanation for prioritization in a form that is accessible to medical professionals. The algorithm is well suited to the growing amount of phenotype-genotype information in databases such as ClinVar.

Keywords: Human Phenotype Ontology, phenotype-driven genomic diagnostics, evidence-based medicine

1 Introduction

Diagnostic exome or genome sequencing usually reveals a large number of variants that are considered deleterious by current computational methods. Therefore, the analysis of such data usually requires an additional criterion to prioritize genes for further investigation.

Phenotypic approaches use the proband's observed phenotypic abnormalities to evaluate disease candidates by searching for diseases with similar phenotypic abnormalities associated with genes harboring a predicted pathogenic variant.

*Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Vrazov trg 2, Ljubljana SI-1000, Slovenia

†Correspondence: matevz@inetis.com

‡Present address: Inetis Ltd, Celje SI-3000, Slovenia

Several heuristic criteria, such as IC [1], likelihood ratio [2] and various phenotype similarity measures have been proposed as a guiding criteria in gene prioritization process.

Heuristics can be useful, but they are still a calculated *guess* and may also contain various biases (e.g., omitted or presupposed evidence, implicit or unsound estimates of various quantities, etc.).

A viable alternative to heuristic methods are evidence-based methods, which are characterized by using only and exclusively the information available to the clinician. Estimates are used as a last resort when data sets are sparse.

In this paper, we first formulate a general problem of medical diagnostics in a Bayesian framework and apply it to the problem of gene prioritization. Following the principles of evidence-based medicine, only training cases and the HPO ontology are used and no additional assumptions or ad-hoc criteria are introduced. The resulting algorithm WA calculates the plausibility of a gene's pathogenicity in a form that allows interpretation and explanation of the gene prioritization results.

The implementation of evidence-based model results in a simple, robust, and efficient algorithm WA which can be implemented in a few lines of code. The algorithm was highly accurate compared to the state-of-the art methods.

2 Material and Methods

2.1 Human Phenotype Ontology

The Human Phenotype Ontology is a standardized and controlled representation of human phenotypic abnormalities. The phenotypic terms in HPO are represented as a directed acyclic graph (DAG). Nodes in DAG represent phenotypic signs and graph arrows represent 'is a' relationship between phenotypic signs (e.g. Rhabdomyoma is a special case of Neoplasm of striated muscle). In our study, we used the HPO version, downloaded via HPO rest API [3], which contains 16 290 human phenotypic terms.

2.2 Datasets

The published CADA implementation includes a dataset of 4 714 cases, divided into test, training, and validation sets [4]. The data were collected partly from their clinical collaborators and partly from the ClinVar database [5]. Because the ClinVar database is variant-based it was decided to merge variants in recessive

genes from the same submitter that are characterized by the same phenotypic features, assuming compound heterozygosity ([6], Sect Clinical cases).

We have decided not to merge ClinVar cases and to treat each variant-proband-phenotype triple classified as pathogenic or likely pathogenic as a separate case. The reason for this is that in Mendelian diseases, a single pathogenic variant in a gene can lead to a gene-related disease. Therefore, a single pathogenic variant can be considered as evidence of a gene-phenotype association. Our ClinVar database extract consists of 40 752 training cases.

For testing only we use a set of 665 cases of probands diagnosed at the Clinical Institute of Genomic Medicine, University Medical Centre Ljubljana (UMCL dataset). The cases are not included in any public database and represent a real-word distribution of cases in the SE Europe region.

2.3 Prioritization Model

To assist the diagnostic process, it is helpful to estimate the probability that a patient has disease D. After collecting evidence $E = E_1, \ldots, E_n$, the theoretically sound procedure is to calculate the probability that the patient has disease D, given all the evidence collected:

$$P(D \mid E) = P(D \mid E_1 E_2 \ldots E_n) \tag{1}$$

and prioritize candidate list of diseases accordingly.

In addition to the evidence collected pertaining to an individual patient (e.g., various test results and the observed phenotype) there is a body of knowledge B that is relevant to D [7]. Clinicians know, for example, the prevalence of the disease in the population, its onset, the pattern of inheritance, etc. Over the years of practice, clinicians may also accumulate tacit knowledge [8], i.e., a 'gut feeling' that is a valid, if not perfect, part of a clinician's background knowledge. To include the background knowledge B, we need to rewrite the Equation 1 as:

$$P(D \mid EB) = P(D \mid E_1 E_2 \ldots E_n B) \tag{2}$$

The inclusion of background knowledge B in 2 is mandatory in order to remain true to the basic principle of evidence-based medicine: all available relevant evidence must be considered, not just an arbitrary subset of it [9]. Disregarding this principle would lead us either to ignore the evidence that is available or to presuppose the evidence that is not available [7].

We now proceed to derive Equation 2 into a form that can be used for prioritizing diseases. We consider the case of estimating the probability of disease D against

the probability of all other diseases $\overline{D}$. Using probability product rule in the form: $P(DE \mid B) = P(E \mid DB)P(D \mid B) = P(D \mid EB)P(E \mid B)$ we first rewrite Equation 2 as:

$$P(D \mid EB) = P(D \mid B)\ \frac{P(E \mid DB)}{P(E \mid B)} \tag{3}$$

Since estimating the probability of evidence $P(E \mid B)$ is difficult, we eliminate it by focusing on the ratio of the probability that patient has a disease E to the probability that the patient does not have a disease E, which is called the odds:

$$O(D \mid EB) = \frac{P(D \mid EB)}{P(\overline{D} \mid EB)} = \frac{P(D \mid B)\ P(E \mid DB)}{P(\overline{D} \mid B)\ P(E \mid \overline{D}B)} \tag{4}$$

combination of Equation 3 with Equation 4 gives the posterior odds for the disease B given evidence E and background knowledge B:

$$O(D \mid EB) = O(D \mid B)\ \frac{P(E \mid DB)}{P(E \mid \overline{D}B)} \tag{5}$$

If we take a single piece of evidence E Equation 5 can already be used to estimate the odds and probability of disease D: Information about the prevalence of the disease in the population can be used to estimate $O(D \mid B) = \frac{P(D \mid B)}{P(\overline{D} \mid B)}$, and background knowledge of the clinician or data from various databases can be used to estimate the probability of evidence $P(E \mid DB)$ (e.g. test result, morphological sign) if we know that the patient has disease D. Similarly, the probability of evidence $P(E \mid \overline{D}B)$ is estimated under the assumption that we knew the patient did not have disease D.

Before extending the estimate of the probability of D to the multiple evidences $E = E_1, \ldots, E_N$, we first introduce a more intuitive unit for measuring the evidence of D. Differences in the probabilities of different diseases that are close to 0 or 1 are difficult to interpret (for example, it is difficult to semantically distinguish the probability of a disease D of .999 from .9999). Good ([10], attributed to Turing) proposed the use of log odds as a measure of the plausibility of a binary hypothesis D given evidence E. The plausibility (evidence) of the disease D, denoted $e(D \mid E)$, is defined as:

$$e(D \mid EB) = 10 \log_{10} O(D \mid EB)\ \ (dB) \tag{6}$$

The definition of plausibility is analogous to the ratio of two sound intensities in acoustics. Since the plausibility of D is defined in terms of logarithms, it is easy to distinguish probabilities of D: 0.999 corresponds to 30 dB, 0.9999 to 40 dB and even odds correspond to 0 dB.

If we rewrite Equation 5 in terms of the plausibility of D, we obtain

$$e(D \mid EB) = e(D \mid B) + 10 \log_{10} \left[\frac{P(E \mid DB)}{P(E \mid \overline{D}B)} \right] \tag{7}$$

To extend Equation 7 to multiple pieces of evidence $E = E_1, \ldots, E_N$ [7] using product rule we get:

$$\begin{aligned} e(D \mid EB) = e(D \mid B) &+ 10 \log_{10} \left[\frac{P(E_1 \mid DB)}{P(E_1 \mid \overline{D}B)} \right] \\ &+ 10 \log_{10} \left[\frac{P(E_2 \mid E_1 DB)}{P(E_2 \mid E_1 \overline{D}B)} \right] \\ &+ 10 \log_{10} \left[\frac{P(E_3 \mid E_2 E_1 DB)}{P(E_3 \mid E_2 E_1 \overline{D}B)} \right] + \cdots \end{aligned} \tag{8}$$

Unfortunately, conditional probabilities $P(E_i|E_{i-1} \ldots E_1 B)$ are rarely available for most medical domains. In such a case, we may decide to assume *logical* independence of pieces of evidence E_i and Equation 8 simplifies to:

$$e(D \mid EB) = e(D \mid B) + 10 \sum_i \log_{10} \left[\frac{P(E_i \mid DB)}{P(E_i \mid \overline{D}B)} \right] \tag{9}$$

2.3.1 Using Prioritization Model

In the case of logically independent evidence E in Equation 9 prioritization is straightforward: after collecting the probabilities $P(E_i \mid DB)$, $P(E_i \mid \overline{D}B)$, and $P(D \mid B)$, the plausibility of disease E is calculated. The same procedure is performed for all diseases. The diseases are ordered by decreasing plausibility, with the first disease being the most probable with respect to evidence E.

In the case of dependent evidence the conditional probabilities in Equation 8 must be estimated. It may well be the case that only a small subset of the evidence is inter-dependent and the rest of the terms in Equation 8 decay into independent terms of Equation 9.

In some cases it may also be interesting to consider the plausibility of *evidence gain* of disease D. The reasons for this are illustrated by the example of a rare disease D_r whose prevalence is well below 1 per million. The term $e(D \mid B)$ in the Equation 8 and 9 can be less than -40 dB. Even if the evidence E increases the plausibility of D_r by 30 dB (i.e. D_r is a thousand times more probable than before), this results in only $P(D_r \mid EB) = 1/11$. D_r may not be high on the priority list, but it may still be worth considering.

2.3.2 Estimation of the Plausibility of Gene Diseases

Even though the Equation 9 is sound, it represents the conditional probability that depends on the conjunction of events $E_1 \wedge E_2 \wedge \cdots$. The probability of the conjunction of events is much smaller than the probability of its conjuncts and therefore very difficult to estimate with a limited number of training cases. To alleviate the problem, we proceed by separately estimating the plausibility of disease D given a single piece of evidence E_i and summing the contributions of the evidence to the final plausibility of disease D. Using the definition of odds in Equation 4 and plausibility in Equation 6 we have:

$$e(D \mid E_i B) = 10 \log_{10} O(D \mid E_i B) \tag{10}$$

$$= 10 \log_{10} \frac{P(D \mid B)\ P(E_i \mid DB)}{P(\overline{D} \mid B)\ P(E_i \mid \overline{D} B)} \tag{11}$$

In the gene prioritization problem, the disease D corresponds to the gene G with a pathogenic variant and the evidence E_i is represented by the proband's phenotype sign ϕ_i. As a side note, to account for the cases of patients without genetic diseases, we can introduce a special genetic label G^* for such cases. We can also omit the background knowledge B, as it is not used in our experiments, which gives the plausibility of a pathogenic variant in gene G given a phenotypic sign ϕ_i:

$$e(G \mid \phi_i) = 10 \log_{10} \frac{P(G)\ P(\phi_i \mid G)}{P(\overline{G})\ P(\phi_i \mid \overline{G})} \tag{12}$$

$$= 10 \log_{10} \frac{P(G)\ P(\phi_i \mid G)}{\sum_{G' \in \overline{G}} P(G')\ P(\phi_i \mid G')} \tag{13}$$

If values of the numerator or denominator in Equation 13 are 0, we set the plausibility to -30 dB and 30 dB, respectively.

The estimate of the plausibility of the pathogenic variant in gene G given the observed phenotype of the proband $\Phi = \{\phi_i\}$ is estimated as the sum of the plausibility over all phenotypic signs:

$$e(G \mid \Phi) = \sum_{\phi_i \in \Phi} e(G \mid \phi_i) \tag{14}$$

$$= 10 \sum_{\phi_i \in \Phi} \log_{10} \frac{P(G)\ P(\phi_i \mid G)}{\sum_{G' \in \overline{G}} P(G')\ P(\phi_i \mid G')} \tag{15}$$

2.4 Extended Phenotype

A phenotype ontology such as HPO provides metadata of phenotypes that can be used in prioritizing genes. Note that each phenotypic sign ϕ_i in the proband's phenotype Φ is part of the HPO ontology, which is represented as an acyclic graph of phenotypical signs. If ϕ_i is observed in a proband, we can *deductively* infer that the proband also has all the phenotypic signs in the ϕ_i ancestor set in the HPO ontology (e.g. if a proband has a rhabdomyoma, we can infer that he also has a neoplasm of striated muscle, an abnormality of musculature, an abnormality of the musculoskeletal system, and a phenotypic abnormality).

If $A(\phi_i), \phi_i \in \Phi$ denotes the set of all ancestors of ϕ_i (including ϕ_i) in the HPO ontology, we define the extended proband phenotype Φ^* as

$$\Phi^* = \bigcup_{\phi_i \in \Phi} A(\phi_i) \tag{16}$$

2.5 WA algorithm

WA is a simple and efficient algorithm for gene prioritization based on the prioritization model in its estimated form (see Sect 2.3.2 and 2.4).

The learning phase of the WA algorithm computes the information for gene prioritization: the probabilities needed for the plausibility calculation are estimated from the test cases provided to the algorithm. WA pseudo code for the learning phase is presented in Algorithm 1. The prioritization phase of WA estimates the plausibility of the pathogenicity of each gene given the proband's phenotype. The list of genes is sorted by plausibility in descending order. In case of a tie (i.e., two genes have the same plausibility score), the genes are sorted according to the prior probability of the gene estimated from the training set. WA pseudocode for the gene prioritization phase is presented in algorithm 2.

2.6 The Interpretation and Explanation of Gene Prioritization

There are many advantages to developing an explainable clinical decision support systems. Understanding why a system has prioritized a particular gene can help convince a clinician that it is legitimate.

The decision model of the WA algorithm can be explained simply: The phenotypic sign ϕ of the proband gives a vote ($e(G \mid \phi)$, equation 13) for each candidate gene G. The gene is prioritized according to the sum of the votes from all proband's phenotypical signs.

Algorithm 1: WA - learning phase

Input: A set of all genes G, a set of all phenotypic signs Φ from the HPO, a list of training cases $\langle G_1, \Phi_1 \rangle, \ldots \langle G_n, \Phi_n \rangle$, where $G_i \in G$ is a gene with pathogenic variation and $\Phi_i \subseteq \Phi$ is a proband phenotype, $A(\cdot)$ is an ancestor function.

Output: Set of probability estimates used in prioritization phase of WA algorithm.

for $g \in G$, $\phi \in \Phi$ **do**
    $\text{freq}[g] \leftarrow 0$
    $\text{freq}[\phi \mid g] \leftarrow 0$
// compute frequencies in the test set
for $i \leftarrow 1$ **to** n **do**
    $\texttt{inc}(\text{freq}[G_i])$
    $\Phi^* \leftarrow \bigcup_{\phi \in \Phi_i} A(\phi)$ // See Equation 16
    for $\phi \in \Phi^*$ **do**
        $\texttt{inc}(\text{freq}[\phi \mid G_i])$
for $g \in G$, $\phi \in \Phi$ **do**
    $P(g) \leftarrow \text{freq}[g] \,/\, n$
    if $\text{freq}[g] > 0$ **then**
        $P(\phi \mid g) \leftarrow \text{freq}[\phi \mid g] \,/\, \text{freq}[g]$
    else
        $P(\phi \mid g) \leftarrow 0$
return $\{P(g), P(\phi \mid g) \mid g \in G, \phi \in \Phi\}$

Algorithm 2: WA - gene prioritization

Input: A set of all genes G, a set of all phenotypic signs Φ, $\Phi_p \subseteq \Phi$ is a proband phenotype, $\{P(g), P(\phi \mid g) \mid g \in G, \phi \in \Phi\}$ is a set of probability estimates from the learning phase of WA algorithm.

Output: list of sorted pairs $\langle\, g, \texttt{plausibility}(g) \,\rangle$ where $g \in G$.

$\Phi^* \leftarrow \bigcup_{\phi \in \Phi_p} A(\phi)$ `// See Equation 16`

$L \leftarrow []$

for $g \in G$ **do**

    $\texttt{append}(\langle\, g, \sum_{\phi \in \Phi^*} \texttt{plausibility}(g, \phi) \,\rangle, L)$

return $\texttt{sort\_descending}(L, \texttt{less})$ `// sort by plausibility`

`// Equation 12:`

Function $\texttt{plausibility}(g,\ \phi)$

    $P(g \wedge \phi) \leftarrow P(\phi \mid g)\, P(g)$

    if $P(g \wedge \phi) = 0$ **then**

        return -30

    $P(\overline{g} \wedge \phi) \leftarrow \sum_{g' \in G - \{g\}} P(\phi \mid g')\, P(g')$

    if $P(\overline{g} \wedge \phi) = 0$ **then**

        return 30

    return $10 \log_{10} \{\, P(g \wedge \phi) \,/\, P(\overline{g} \wedge \phi) \,\}$

Function $\texttt{less}(\langle\, g_1, \texttt{plausibility}(g_1) \,\rangle, \langle\, g_2, \texttt{plausibility}(g_2) \,\rangle)$

    if $\texttt{plausibility}(g_1) \neq \texttt{plausibility}(g_2)$ **then**

        return $\texttt{plausibility}(g_1) < \texttt{plausibility}(g_2)$

    else

        return $P(g_1) < P(g_2)$

The voting and propensity of each phenotypical sign ϕ towards a gene G can be visually represented in the form of a heatmap table. One column of the heatmap table represents the distribution of votes for the phenotypic sign ϕ of a proband for all prioritized genes. One row of the heatmap table represents all votes of the phenotypic signs of the proband for a specific gene.

Table 1 represent result of a prioritization for a test case form the CADA database. From the distribution of columns we can easily observe a strong propensity of phenotypical signs 'Cherry red spot of the macula' and 'Hyperacusis' towards *HEXA* gene. The phenotypical sign 'Developmental regression' and 'seizure' in the first and the last column of the table, respectively, spread their votes much more evenly. This is not surprising since developmental regression and seizures are quite common among rare disease patients. If needed the heatmap can be extended to show the exact distribution of the training cases regarding every gene-phenotypical sign combination from every cell of the heatmap.

Table 1 represents the result of a prioritization for a test case from the CADA database From the distribution of columns, it is easy to see a strong tendency of the phenotypic signs 'Cherry red spot of the macula' and 'Hyperacusis' to the gene *HEXA*. The phenotypic sign 'Developmental regression' in the first column of the table distributes its votes much more evenly. This is not surprising, since developmental regressions are quite common in patients with rare diseases. If needed, the heatmap can be expanded to show the exact distribution of training cases with respect to each combination of gene and phenotypic trait in each cell of the heatmap.

In the case of Table 1 the *HEXA* gene clearly dominates over all phenotypical signs except 'Hepatomegaly' and is indeed correctly prioritized by WA. It should be noted that there is generally no guarantee that a single gene is dominant, and in such cases heatmaps provide useful guidance and support for the diagnostic process.

3 Results

In evaluating the quality of the algorithm WA, we focus primarily on prioritization accuracy compared to modern gene prioritization methods CADA [6] and LIRICAL [2]. The effect of expanding the training set of the algorithm on classification accuracy is also measured.

The CADA tests were performed using the online version of the algorithm [11]. The LIRICAL tests were performed using LIRICAL ver v2.0.0-RC1 [12].

Table 1: WA heatmap table for CADA test case (all values in dB)

HP:0002376 Developmental regression	HP:0002353 EEG abnormality	HP:0002240 Hepatomegaly	HP:0010780 Hyperacusis	HP:0010729 Cherry red spot of the macula	HP:0001250 Seizure	
−6.8	−12.1	−14.9	4	4.4	−17.3	*HEXA*
−18.4	−30	−14.9	−7.8	−4.4	−30	*GLB1*
−8.9	−12.1	−30	−30	−30	−15.1	*MECP2*
−15.3	−12.1	−30	−30	−30	−18.5	*GNB1*
−12.2	−30	−4.8	−30	−30	−30	*GNPTAB*
−13.5	−15.3	−30	−30	−30	−19.3	*TPP1*
−15.3	−13.4	−30	−30	−30	−20.3	*SCN8A*
−15.3	−15.3	−30	−30	−30	−20.3	*DEAF1*
−13.5	−15.3	−30	−30	−30	−26.4	*IRF2BPL*
−30	−12.1	−30	−30	−30	−15.8	*ZEB2*
−14.9	−16.7	−24.5	−24.4	−24	−21.3	mean

Since LIRICAL prioritizes diseases, its prioritization results were mapped to the corresponding genes.

3.1 Testing on CADA Data

First we compare the algorithms with the CADA datasets [4]. The algorithm WA does not require validation data, so we train it on both CADA's training and validation datasets. CADA's prioritization accuracy refers to its best results (i.e.

all validation data included in the training phase of the CADA algorithm) and LIRICAL's results refer to its default parametrization. WA outperforms CADA on the top 1, 5, and 10 prioritized genes, and lags behind CADA on the other measured ranks. LIRICAL's results were worse compared to WA and CADA at all ranks.

When WA uses its own version of the ClinVar data with the CADA training cases removed, it outperforms CADA on all measured ranks (Table 2) except the last one.

Table 2: Performance comparison of gene prioritization methods in 943 CADA case (all values in %, the best result for every rank emphasized)

rank %	WA[a]	WA[b]	CADA	LIRICAL
top 1	**29.9**	26.2	19.3	15.8
top 5	**44.2**	38.1	34.6	26.8
top 10	**48.6**	41.4	41.2	33.9
top 50	**61.7**	53.0	58.0	50.0
top 100	66.2	58.4	**67.7**	60.2

[a] ClinVar without CADA test cases + train CADA cases + validation CADA cases used for learning.

[b] train CADA cases + validation CADA cases used for learning.

3.2 Testing on UMCL Data

The UMCL dataset is interesting because it is not part of a public database and can be considered novel for all algorithms.

As can be seen from Table 3, the prioritization accuracy of CADA decreases across all classified ranks compared to testing on the CADA test cases (compare the results in Tables 2 and 3). The accuracy of LIRICAL decreases only on ranks 1, 5 and 10.

The prioritization accuracy of WA trained on CADA datasets decreases for the top rank and remains comparable to testing on CADA test cases for all other ranks.

The prioritization accuracy of WA, which was trained on ClinVar data, improves compared to the accuracy of WA which was trained on CADA training cases only. The results show that the prioritization accuracy of WA remains stable on a new

dataset and that the assumptions regarding the transformation of ClinVar data to training cases (see Sect2.2 section) and evidence estimation (see Sect 2.3.2) lead to good performance on all test cases.

Table 3: Performance comparison of gene prioritization accuracy in 649 UMCL cases (all values in %, the best result for every rank emphasized)

rank %	WA[a]	WA[b]	CADA	LIRICAL
top 1	**21.20**	16.90	4.14	6.61
top 5	**39.63**	36.10	9.98	20.28
top 10	**47.93**	42.24	16.43	30.41
top 50	**60.22**	54.07	39.01	52.84
top 100	**65.13**	57.60	50.38	64.21

[a] ClinVar + all CADA cases used for learning.
[b] all CADA cases used for learning

4 Discussion

Clinical decision support systems are important tools that help clinicians in genomic diagnostics. In addition to the accuracy of their results, the interpretation and explanation of their results in the clinical setting is also critical.

Most tools used for gene prioritization are based on heuristic criteria, chosen primarily for their accuracy on available training cases. We decided to start with a general definition of the problem of evidence-based medical diagnosis and formulate it in a Bayesian framework. Evidence-based medical diagnostics is first stated as the most general question - what is the probability of a disease given all available evidence and background knowledge (see Equation 1). The answer is first derived in its most general form and only then adapted to the gene prioritization problem. In the final stage, probability estimates for various quantities are introduced. The advantage of formulating gene prioritization within a general framework is that all differences between the theoretically optimal decision model and the adaptations introduced due to the lack of information or training cases are made transparent.

The final model for gene prioritization results in a simple, efficient algorithm with accuracy that is significantly better than state-of-the-art approaches on existing and new datasets. In addition, the model provides clear and intuitive explanations

for its decisions. The decision process can be interpreted to mean that proband's phenotypes vote for genes and the gene voted for the most is the most likely cause of the disease.

The plausibility of a disease given evidence (see Equation 10) and its adaptation for the problem of gene prioritization without using background knowledge (see Equation 12) not only includes the likelihood ratio [2], but also takes into account the prior probability of a disease estimated from the training dataset. The formulation in decibel units can be interpreted as the ratio of the pressure levels of two sounds, one of which is the sound of the disease with a phenotypic sign in the proband's phenotype and the other is the sum of the sounds of all other diseases with the same phenotypic sign.

The contribution of the proband's phenotypic signs may also be useful in clinical practice. A column in the heatmap (see Table 1) represents the distribution of votes of a patient's phenotypic sign for genes. For phenotypic signs shared by many rare diseases (e.g. seizure, developmental retardation) or for relatively common phenotypic signs in the general population, we can expect an even distribution of votes; however, for specific symptoms, phenotype's votes are much more concentrated among good candidates for disease-causing genes. Looking at the columns of the heat map can also help a clinician distinguish disease symptoms from incidental phenotypic signs. The resulting gene list in the heatmap may even lead the clinician to reconsider typically incidental symptoms in light of additional information (e.g., toddler with high myopia [2]).

WA algorithm does not introduce additional assumptions regarding the proband's phenotype (e.g., phenotype similarity, which is often difficult to estimate in sparse datasets). The extended phenotype of the proband does not require probability estimates because its construction is based only on logically inferable information (i.e., it includes only the ancestors of the phenotype with respect to the HPO ontology).

Finally, a word about the structure of the gene prioritization algorithm: WA uses only two concepts (test case represented as the set of proband's phenotypic signs associated with the pathogenic gene and the HPO ontology) and a minimal set of assumptions (most importantly, logical independence of patient phenotypic signs, probability estimation of events with relative frequencies, and pruning of plausibility values in the case of undefined values of Equation 13). Structurally simple systems usually have the advantage of being easy to analyze, change or extend. WA has already built in an option for very general extensions with the introduction of background knowledge. Despite its simplicity, results suggest that WA shows superior performance on real-world datasets.

4.1 Limitations

The Bayesian diagnostic model is complete in the sense that it considers all and only the information relevant to the diagnosis. However, in its fully general form, the conditional probability of a disease given all the circumstantial evidence and background knowledge (Equation 2) is difficult to estimate because it involves the conjunction of all the circumstantial evidence. WA assumes that the proband's phenotypic signs (i.e., the available evidence) are logically independent, which is called 'naive Bayes'. This assumption is unrealistic, but studies show [13] that naive Bayes works well for feature distributions with low entropy, and our results show the same.

Although a fully general diagnostic model is probably not achievable with the current set of available training cases, a feasible extension of WA would be to include information on pairs of phenotypic features, as suggested by [2]. This could be done in a principled way, either by introducing a new evidence label for each pair of phenotypic features that can be reliably estimated from the training cases, or by including the conditional probability of phenotypic pairs in all equations derived from Equation 9.

Background knowledge of a clinician (e.g., severe myopia in a toddler must be considered differently than in an adult) and general medical background knowledge (e.g., type of inheritance of a disease, mandatory phenotypic signs present or absent in a disease, etc.) are also not considered when implementing WA. Again, background knowledge can be accounted for in a principled manner by including it in all equations derived from Equation 9.

Finally, variant pathogenicity information from the ClinVar database can be included if the proband's genomic information is available. In this case, no extension of the model or the WA algorithm is required, as variant information is treated in the same way as any other type of evidence (e.g., phenotypic signs in the gene prioritization problem).

Data and Code Availability

The CADA data used in this paper and the implementation of the WA algorithm can be found at https://github.com/matevz-kovacic/WA.

Acknowledgments

Many thanks to the Clinical Institute of Genomic Medicine at University Medical Centre Ljubljana for granting access to the anonymized UMCL dataset.

I also thank Gaber Bergant for his invaluable information on the problem of gene prioritization.

This research was funded by the Slovenian Research Agency (grant number J5-1780 "Using Literature-based Discovery for Interpretation of Next Generation Sequencing Results"). The funding source had no role in study design, data collection and analysis, interpretation of data, decision to publish, or preparation of the manuscript.

Declaration of Interests

The author declares no competing interests.

References

[1] Sebastian Köhler, Marcel H. Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E. Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N. Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457–464, 2009.

[2] Peter N Robinson, Vida Ravanmehr, Julius O B Jacobsen, Daniel Danis, Xingmin Aaron Zhang, Leigh C Carmody, Michael A Gargano, Courtney L Thaxton, Guy Karlebach, Justin Reese, Manuel Holtgrewe, Sebastian Köhler, Julie A McMurry, Melissa A Haendel, and Damian Smedley. Interpretable clinical genomics with a likelihood ratio paradigm. *The American Journal of Human Genetics*, 107(3):403–417, 09 2020.

[3] The Jackson Laboratory. Hpo ontology. `https://hpo.jax.org/webjars/swagger-ui/3.20.9/index.html?url=/api/hpo/docs`, 2021. [Online; accessed 25-Dec-2021].

[4] Chengyao Peng. Cada datasets. `https://github.com/Chengyao-Peng/CADA/tree/main/data/processed/cases`, 2021. [Online; accessed 09-Aug-2022].

[5] National Institutes of Health (NIH). Clinvar dataset. https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz, 2022. [Online; accessed 20-Oct-2022].

[6] Chengyao Peng, Simon Dieck, Alexander Schmid, Ashar Ahmad, Alexej Knaus, Maren Wenzel, Laura Mehnert, Birgit Zirn, Tobias Haack, Stephan Ossowski, Matias Wagner, Theresa Brunet, Nadja Ehmke, Magdalena Danyel, Stanislav Rosnev, Tom Kamphans, Guy Nadav, Nicole Fleischer, Holger Fröhlich, and Peter Krawitz. CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. *NAR Genomics and Bioinformatics*, 3(3), 09 2021.

[7] E. T. Jaynes. *Probability theory: The logic of science*, pages 86–90. Cambridge University Press, 2003.

[8] M. Polanyi. *Personal Knowledge: Towards a Post-Critical Philosophy*. University of Chicago Press, 1958.

[9] D.L. Sackett, W.M.C. Rosenberg, J.A. Gray, Haynes, and W.S. Richardson. Evidence based medicine: What it is and what it isn't. *BMJ (Clinical research ed.)*, 312:71–2, 02 1996.

[10] I.J. Good. *Probability and the Weighing of Evidence*, pages 62–64. Charles Griffin, 1950.

[11] GeneTalk GmbH. Cada.webservice. https://cada.gene-talk.de/webservice/, 2022.

[12] Lirical code. https://github.com/TheJacksonLaboratory/LIRICAL/releases, 2022. [Online; accessed 02-Nov-2022].

[13] Thomas J. Watson. An empirical study of the naive bayes classifier. 2001.