bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Klarigi: Characteristic Explanations for Semantic Data

Luke T. Slater[1,2,6,8*], John A. Williams[1,2], Paul N. Schofield[9], Sophie Russell[1,2,6,8], Samantha C. Pendleton[1,2,6,8], Andreas Karwath[1,2,6,8], Hilary Fanning[2,8], Simon Ball[2,6,8], Robert Hoehndorf[7], Georgios V Gkoutos[1,2,3,4,5,6,8]

1 College of Medical and Dental Sciences, Institute of Cancer and Genomic Sciences, University of Birmingham
2 Institute of Translational Medicine, University Hospitals Birmingham, NHS Foundation Trust
3 NIHR Experimental Cancer Medicine Centre
4 NIHR Surgical Reconstruction and Microbiology Research Centre
5 NIHR Biomedical Research Centre
6 MRC Health Data Research UK (HDR UK) Midlands
7 Computational Bioscience Research Center, Computer, Electrical and Mathematical Sciences & Engineering Division, King Abdullah University of Science and Technology
8 University Hospitals Birmingham NHS Foundation Trust, Edgbaston, Birmingham, UK
9 Dept of Physiology, Development, and Neuroscience, University of Cambridge, UK

*** Corresponding author: l.slater.1@bham.ac.uk**
Keywords: ontology, semantic analysis, enrichment analysis, semantic explanation, phenotypes

Abstract

Background Annotation of biomedical entities with ontology terms facilitates the use of background knowledge in analysis. Described entities may be stratified into groups or otherwise assigned labels, and it is of interest to identify semantic characterisations of these groups based on their ontological annotations. Enrichment analysis is routinely employed to identify classes that are over-represented in annotations across sets of groups, most often applied to gene set analysis. However, these approaches usually consider only univariate relationships, make limited use of the semantic features of ontologies, and provide limited information and evaluation of the explanatory power of both singular and grouped candidate classes. Moreover, they do not solve the problem of deriving cohesive, characteristic, and discriminatory sets of terms for entity groups.

Results We develop a new tool, Klarigi, which introduces multiple scoring heuristics used to identify classes that are both explanatory and discriminatory for groups of entities annotated with ontology terms. The tool includes a novel algorithm for derivation of multivariable semantic explanations for entity groups, makes use of semantic inference through live use of an ontology reasoner, and includes a classification method for identifying the discriminatory power of candidate sets. We describe the design and implementation of Klarigi, and evaluate its use in two test cases, comparing and contrasting methods and results with literature and enrichment analysis methods.

Conclusions We demonstrate that Klarigi produces characteristic and discriminatory explanations for groups of biomedical entities in two settings. We also show that these explanations recapitulate and extend the knowledge held in existing biomedical databases and literature for several diseases. As such, Klarigi provides a distinct and valuable perspective on biomedical datasets when compared with traditional enrichment methods, and therefore constitutes a new method by which biomedical datasets can be explored, contributing to improved insight into semantic data.

Background

Over the last two decades, it has become standard practice to represent biomedical knowledge using ontologies. Ontologies define controlled domain representations for biomedical concepts, and the linkage of data to these representations is intended to reduce variability and ambiguity, as well as to synthesise additional knowledge through contextual aggregation. As such, biomedical ontologies and their instance data constitute a massive source of formalised knowledge.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

This repository of formalised knowledge facilitates secondary use, taking advantage of the semantic features of ontologies to integrate and explicate upon the large and multi-modal datasets annotated to them. Semantic analysis methods are applied to many knowledge synthesis and classification tasks, including prediction of protein interaction and rare disease variants [1]. In the clinical space, semantic methods have been applied to tasks including diagnosis of rare and common diseases [2, 3] and identification of subtypes of diseases such as autism [4]. In addition, the synthesis of classical ontology-based methods and machine learning is increasingly common [5].

Despite the increasing use of semantics in biomedical analysis, the subsequent derivation of semantic explanations for classifications, outcomes, labels, or groups generated by those analyses, remains a challenging task, and a major practical, ethical, and technical issue.

We define the task of semantic explanation as that of producing, given a set of entities described by ontology terms, a set of terms that characterises the set of entities. For example, and most prominently, gene set enrichment analysis (GSEA) identifies terms that are significantly over-expressed in a set of genes [6]. Enrichment analysis techniques have also been used to explore over-representation of terms in groups of entities associated with other datasets and ontologies [7].

There are, however, multiple drawbacks to the use of enrichment analysis methods for semantic explanation:

- Traditional enrichment methods using a univariate statistical test provide limited information about the relationship of each explanatory term with the group of interest.
- They usually do not consider the multivariate contribution of sets of terms to a non-redundant characterisation of the group of interest.
- Most make little or no use of the semantic features of ontologies, or information-theoretic methods of evaluating candidate terms.
- They provide limited tools and metrics for the evaluation of the overall explanatory power of groups of terms with respect to the group of interest.
- They can be difficult or impossible to use with non-preconfigured ontologies, often with a dedicated focus on the Gene Ontology (GO) [8].

Moreover, enrichment methods do not attempt to identify a full and cohesive set of characteristic terms for a group, and these methods therefore provide a limited solution to the task of semantic explanation. We propose to solve these problems by developing a new method for semantic explanation, with the aim of deriving characteristic and discriminatory sets of explanatory terms for groups. The method introduces new metrics for measurement for contribution of both individual and groups of candidate classes to the characterisation of a group. The new metrics include a configurable measure of information content, facilitating the use of the information theoretic, in addition to the purely statistical, mode of identifying explanatory terms. We also introduce a new algorithm for derivation of multivariable sets of terms that together provide a characteristic explanation of a maximal subset of the group of interest, within configurable boundaries. The method also makes use of ontological inference, by live use of the ELK reasoner [9]. Furthermore, the method is ontology agnostic, and can be used with any OWL ontology without any pre-processing. To provide tools for evaluating the characteristic and discriminatory power of explanatory sets produced, we also introduce a new classification method that can be used to reclassify the evaluated data, as well as to classify a holdout test set or other new data.

In this work, we formally describe the method, including the metrics and algorithms involved. In doing this, we discuss existing approaches and their potential limitations, explaining how our proposed approach solves these problems. We then describe development of the tool, Klarigi, that implements the method. The work is inspired by a previous investigation, in which we developed an algorithm for deriving multivariable explanations for semantic clusters identified from Human Phenotype Ontology (HPO) phenotype profiles [10]. In this work, the approach is heavily modified and improved upon, including changes to the algorithm and scoring system, and is generalised to be applicable to any dataset and ontology. We also include native support for the Phenopacket format [11]. The resulting tool provides an analysis framework that can be used to explore any dataset annotated by a consistent Web Ontology Language (OWL) ontology.

To evaluate our approach, we develop two use cases for characterisation of diseases. We compare and contrast the results of the analysis with those of an enrichment analysis, as well as existing literature and biomedical databases. In

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

doing this, we show that Klarigi can be used to provide insights into biomedical datasets via semantic explanation, providing a perspective that is both distinct and complementary to the use of enrichment.

Literature Review

There are numerous methods developed for the R statistical programming environment, as well as online applications, to assess gene over-expression. Most of these include built-in or API calls to annotations to a single ontology: fgsea [12] and GSEABase [13] leverage the Gene Ontology [14] exclusively; DOSE [15] implements the Disease Ontology [16]; ReactomePA [17] makes use of the Reactome pathway database; MamPhea [18] encodes the Mammalian Phenotype Ontology [19]. Additional methods implement multiple ontologies as packages or online tools, including DAVID [20], GProfiler [21], and XGR [22]. Of these, only XGR allows users to provide their own custom ontology, and this must be a graph object mutated from an OWL file. Additionally the XGR package is the only one of these which does not explicitly rely on prepopulated, species-specific gene-ontology annotation databases. This limits analyses to both species-specific questions and to genes or gene products, and these tools have little direct relevance to non-genetic translational medicine.

An additional limiting factor in these analyses is their univariate approach: each ontology entity is tested for overrepresentation on a univariate basis. While methods do exist to help correct for the dependence of statistical tests on related ontology entities, these are post-hoc corrections which do not fully take advantage of the information gained by using an formal ontology. While XGR and Func [23] both take the ontology structure into account, both require manipulation of the well-used OWL file format beforehand if using ontologies other than the Gene Ontology (Func) or using ontologies outside a predefined set or not using pre-computed gene/ontology annotation databases (XGR). Recent Bayesian approaches to ontology enrichment likewise use ontology structure, but require additional numerical data (effect sizes, variance measures from experiments) in to construct informative priors [24] or are restricted to the Gene Ontology [25].

Some previous works have explored multivariate gene set enrichment. The tmod R package uses an approach that finds associations between feature sets and significant components derived through principal component analysis [26] which requires raw gene expression values, and makes interpretation difficult due to pre-processing into principal components. One more recent work [27] used a random walk approach, and involved multi-omics data. Traditional GSEA, ranking genes in order of relevance of biological assays and defining a function as the number of genes above that sorted list, can be formulated as a random walk along sorted features. This approach implements multiple sorted lists from multiomics experiments and implements joint random walks among assays. Another recent development uses Bayesian regression to model the relationship between phenotypes and the proportion of patients with rare alleles, and interpreting the probability of associations being equal or depending on the phenotypic similarity of the patients with rare alleles [28]. This approach requires informed assumptions about existing gene/phenotype relationships in literature to weight priors. All of these methods are specific to the genetic domain, and cannot therefore be easily applied to other kinds of biomedical data.

One previously described approach, OntoFunc [29], provides a solution for the use of inferences made by an ontology reasoner in enrichment analysis. It does this by pre-processing a given OWL file, classifying it, and then saving the inferred axioms into a graph that can then be used by other enrichment methods.

Design and Implementation

The fundamental challenge of semantic explanation is to determine, given a set of groups, and sets of entities described by ontology terms associated with those groups, a set of characteristic terms for the group. Since we developed this approach to be applicable to any biomedical ontology, we will describe the approach, including a data model and algorithms, in abstract terms first. We will then describe their implementation into the Klarigi tool.

Data Modelling

Klarigi aims to determine sets of characteristic ontology classes, given a corpus of entities described by ontology terms. Doing this requires, at minimum, a corpus with three features:

1. A set of entities

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2. A set of groups, which are each associated with a set of entities
3. An OWL ontology describing all terms in the corpus

We can define these as sets, which will allow us to define metrics and heuristics on their contents. G is a set of groups, where G_i is the set of entities ascribed to that group. E is a set of entities, where E_j is the set of terms associated with with that entity. O is a set of ontology terms in the ontology, where ever O_x is a term in the ontology.

Metrics

To identify good sets of explanatory classes for groups, we need to balance multiple qualities of the relationship between candidate classes and the group. Specifically, we want to identify classes that are, in comparison with other candidates:

1. Associated with greater proportion of entities in the considered group.
2. More associated with entities in the considered group than in other groups.
3. More informative.

To measure these qualities and thus make them comparable, we define three univariate scores to measure them, encoding numeric measures of the explanatory power of each candidate class for the given group. We first define a simple function that enables us to determine whether or not a particular entity in the corpus is annotated with a class (or any of its subclasses):

$$member(C, N) = \begin{cases} 1 & |C \cap N| > 0, \\ 0 & otherwise \end{cases} \tag{1}$$

Given two sets C and N, any two sets of ontology classes, this function returns 1, if the size of the intersection of those sets is at least 1, and 0 otherwise. This allows us to identify whether any classes associated with a particular entity include our candidate class. Using this function, we can then define the first score, inclusion:

$$inclusion(O_x, G_j) = \frac{\sum_{\substack{z=1 \\ z \in G_j}}^{|E|} member(subclass(O_x), E_z)}{|G_j|} \tag{2}$$

This score measures the proportion of entities in a given group that are annotated with either the candidate term or any of its subclasses. The subclass function is a call to the ontology reasoner, and returns a set of all transitive subclasses of the class passed as argument, including the original class. This enables the account of the deductive inferences made by the ontology reasoner, entailed by the axioms asserted in the considered ontology.

We next measure how uniquely a candidate class characterises the considered group, or its over-representation in the group with respect to others. Our previous work [10] defined a measure of exclusion as:

$$exclusion_{old}(O_i, G_j) = 1 - \frac{\sum_{\substack{x=1 \\ x \neq j}}^{|G|} \sum_{\substack{y=1 \\ y \in G_x}}^{|E|} member(subclass(O_i), E_y)}{\sum_{\substack{x=1 \\ x \neq j}}^{|G|} |G_x|} \tag{3}$$

This provides a measure of the proportion of entities in the dataset that were associated with the class, but are attributed to *other* groups: the fewer entities external to the considered group that are annotated with the class, the greater the exclusion score. This measure, while useful in the context of previous experiments, it has certain drawbacks. In particular for the cases, where group sizes are unbalanced, the $exclusion_{old}$ measure may give rise to figures that are not easily comparable between different groups in the same dataset. For example, in the case of a dataset with 700 pneumonia patients and 300 pulmonary embolism patients, an exclusion score of 0.7 would be assigned with phenotype association with pneumonia, and 0.3 for pulmonary embolism, even if the inclusion was 0.5

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for both of them; this is correct as a measure of absolute risk for the phenotype occurring in dataset, but is not necessarily helpful for exploring which phenotypes are more characteristic of the considered groups, in proportion to each other.

In GSEA, measurement of comparative expression is usually achieved by an *enrichment score*, which is usually defined by z-scores and/or odds and relative risk ratios, that measures the strength of association between the class and group membership; the ratio of B given the presence of A, and its symmetric inverse. However, their interpretation is challenging due to the behaviour of ratios and confusion between odds, relative, and absolute risk [30]. In the above example, the relative risk of both pneumonia and pulmonary embolism would be 1. However, this can become misleading in the presence of larger class imbalances [31]. For example, if we had 950 cases of pneumonia, and 50 cases of pulmonary embolism. If a phenotype appears thrice in this dataset, and two of those are associated with pulmonary embolism, we would get the extremely large values of 37.96 and 39.54 for relative risk and odds ratio respectively. While this is a valid association, it does not take into account the absolute risk of the event occurring with respect to the disparity in the group sizes, and so large values are given for uncommon and unlikely phenotypes, which can be easily misinterpreted. To combat this issue, we define exclusion as:

$$exclusion(O_i, G_j) = \frac{\sum_{\substack{z=1 \\ z \in G_j}}^{|E|} member(subclass(O_x), E_z)}{\sum_{z=1}^{|E|} member(subclass(O_x), E_z)} - \frac{|G_j|}{|E|} \tag{4}$$

In this measure, the proportion of total entities in the corpus that are annotated with the considered class is determined. From this figure, the proportion of the total entities in the dataset that are associated with the current group is subtracted. This measures the overall representation of the class in the group versus the expression of the group in the overall population and can be explained in terms intuitively similar to the calculation of an odds ratio; the odds of the event occurring in the exposed group minus the odds of exposure.

As such, it caters the measurement of the representation of terms in a considered group, balanced with the absolute likelihood of the group appearing in the whole dataset. In the example above, the exclusion score for the phenotype with 0.5 inclusion with respect to both pneumonia and pulmonary embolism would be 0. This is because neither are more exclusively associated with the phenotype when their relative frequencies are taken into account.

In the latter example, our proposed phenotype association profile would lead to pulmonary embolism receiving an exclusion score of 0.617, and pneumonia an exclusion score of -0.61. These scores express the exclusion of the phenotype association with a group in the context of the overall appearance of the group in the dataset.

We further address the problem of balancing the internal and external characteristic power of candidate terms by introducing an additional heuristic to provide a balanced measure of our previously defined inclusion and exclusion scores. We note that the inclusion score is equivalent to the 'absolute risk' of the class being associated with an entity in the given group. On the basis of those two scores, we can further define:

$$r\text{-}score(O_x, G_j) = 2\frac{inclusion(O_x, G_j) * (1 - exclusion(O_x, G_j))}{inclusion(O_x, G_j) + (1 - exclusion(O_x, G_j))} \tag{5}$$

In the case of our latter example with a large class imbalance, our inclusion score for pulmonary embolism is 0.04, while it is 0.001 for pneumonia. Our *r-score*s would then be 0.036 and 0.0002, respectively. This provides a balanced metric between the internal representation of the class in the group, and its external exclusivity in the context of the larger dataset. Since both figures are taken into account, the r-score metric avoids extreme and potentially misleading values of z-score and odds ratio, when used to identify characteristic classes.

In the case that the program is being used with a corpus that only describes a single class, the *r-score* measure is alternatively defined as equivalent to *inclusion* (note that the program can be forced into this mode, see 'exclusive group loading' later). In this case, the algorithm does not consider exclusion at all, and only provides solutions in the context of *inclusion* – describing the internal composition of the group.

To test our third quality, we introduce a measurement of how informative or specific a candidate class is. Since this score does not rely on any relationship between groups and entities, it is calculated only once per ontology term:

$$specificity(O_x) = IC(O_x) \tag{6}$$

Specificity is defined as the result of a particular information content measure. These are configurable, and provided through implementation of the Semantic Measures Library [32], which implements many different

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

information content measures. Currently, we implemented the Zhou [33] and Resnik [34] measures, which are defined as:

$$IC_{Resnik}(x) = -\log(\frac{|I(x)|}{|I(\top)|}) \qquad IC_{Zhou}(x) = \frac{\log(deep(x))}{\log(deep_{max})} \tag{7}$$

The Resnik measure is defined as the reverse log probability of the term appearing in the corpus, and will therefore be greater for classes that are annotated to entities more infrequently. The Zhou method is a measure of how deep the given term is when representing the ontology as a directed acyclic graph, calculated as a proportion of its maximum depth. Therefore, terms that are deeper in the ontology will have more intermediate subsuming classes between it and *Thing* and as a result will receive greater values. The choice of which information content measure to use is given to the user, and the best choice may differ depending on the dataset, ontology, or intended results [35].

Scoring

Now that we have defined a number of measures for the explanatory power of candidate classes, the next steps depend upon identifying and scoring candidate classes in the relevant biomedical ontology. At this stage, we load and classify the ontology, verifying internal consistency and reflecting any structural inferences. Klarigi uses the ELK reasoner to perform this task [9], which supports the EL subset of OWL. Since we sought to support a maximal number of ontologies, datasets, and settings, we chose ELK since it ensures maximum polynomial time classification with respect to the number of axioms in the ontology. Since Klarigi is implemented using OWLAPI [36], however, use with more expressive reasoners can be easily implemented, or ontologies can be pre-processed before passing to Klarigi.

While scores can be derived, theoretically, for all classes in the biomedical ontology, in practice this is not necessary. This is because only classes that subsume classes appearing in our corpus can receive inclusion and exclusion scores above zero. We can therefore define a new set of candidate terms C, the set of candidate terms in the ontology that we will score, as a subset of O:

$$C = \{x \in O \mid \exists_{z \in E} member(subclass(O_x), E_z)\} \tag{8}$$

In practice, this is implemented more efficiently by iterating directly all classes appearing explicitly in the annotation corpus and identifying their transitive superclasses. Once C is determined, the scores are calculated as above for each combination of group and candidate class, excepting for specificity where the score is calculated only once per class.

Candidate Restriction and Univariate Analysis

We have now created the set of C candidate explanatory classes, which consists of the set of all classes that either directly annotate, or subsume classes that directly annotate, all entities in the corpus. The next step is to identify characteristic candidate variables for each group of interest. We do this by creating a new set U_j for each group:

$$\begin{aligned} U_j = \{x \in C \mid & \textit{r-score}(O_x, G_j) >= \textit{min-r-score} \wedge \textit{inclusion}(O_x, G_j) >= \textit{min-inclusion} \\ & \wedge \textit{exclusion}(O_x, G_j) >= \textit{min-exclusion} \wedge specificity(O_x) >= \textit{min-ic}\} \end{aligned} \tag{9}$$

This new subset U_j can be determined for each group, and contains a subset of the C candidate classes that meet the set of minimum restrictions for each score in relation to the considered group: r-score, inclusion, exclusion, and specificity. The process up to this stage consists the univariate mode of operation for Klarigi. At this point, the set U_j can be output for a group of interest, and subsequently interpreted. The minimum cut-offs for scores are controlled by configurable parameters, described in Table 1.

Multivariable Explanatory Sets

In addition to the univariate mode of operation, we also determine an additional method to identify sets of candidate terms that together characterise the group of interest. To do this, we can consider a solution space S, where S_m is any potential subset of U_j, the set of candidate classes that meet the minimum cut-offs for the univariate scores. Our

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Names, descriptions, and default parameters for candidate class restrictions. These parameters define minimum values for the scoring heuristics, which restrict the set of candidate variables that will appear in the univariate analysis output, and be considered in the multivariable analysis stage.

Parameter	Default	Description
min-inclusion	0	Candidate terms with inclusion values below this level won't appear in, or contribute to, explanatory sets.
min-exclusion	0.05	Candidate terms with exclusion values below this level won't appear in, or contribute to, explanatory sets.
min-r-score	0.1	Candidate terms with *r-score* values below this level won't appear in, or contribute to, explanatory sets.
min-ic	0.4	Candidate terms with information content values below this level won't appear in, or contribute to, explanatory sets.

goal is to identify such an S_m that is a good characterisation of entities in the group. To identify what constitutes a good characterisation of the group, we turn to our previously defined scores. For the purposes of our solution, we only consider *r-score* and $specificity$ to evaluate the explanatory power of individual terms, relying on the ability of *r-score* to balance between *inclusion* and *exclusion* scores. To evaluate the overall fitness of a set of terms S_m, we define an additional heuristic:

$$overall_inclusion(S_c, G_j) = \frac{\sum_{\substack{z=1 \\ z \in G_j}}^{|E|} member(subclass(S_c), E_z)}{|G_j|} \tag{10}$$

This is a measure of overall inclusion, the proportion of entities in the group of interest that are annotated by at least one of the classes in the proposed solution, or their subclasses. This measures how well a set of candidate classes covers the full set of entities in the group.

The challenge of obtaining a good solution lies with the optimisation of several variables, and can therefore can be considered as a multiple objective optimisation problem, considering the scoring heuristics as objective functions. The objective functions can be defined as the cut-offs for our scores. In the case of scores that measure the explanatory power of individual terms, this is the cut-off for candidate terms to appear in S, while the *overall_inclusion* function can be defined as the cut-off for S to be considered acceptable.

This can be considered in terms of the ε-constraints solution to multiple objective optimisation problems. In these solutions, one objective function is retained, while the rest are transformed into a set of constraints between which the remaining objective function can be optimised [37]. Our solution is inspired by this approach, selecting *overall_inclusion* as the objective function. However, instead of using static constraints for the other cutoffs, we develop an algorithm that steps down through acceptable values of these constraints in a priority order to dynamically identify high values for constraints in the context of objective function optimisation.

To do this, we define an order of priority for parameters:

1. *overall_inclusion*
2. *r-score*
3. $specificity$

The objective function can be considered as the highest priority constraint (and is also one that has no lower boundary). To identify a solution that maximises *overall_inclusion*, while optimising the other values within their

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

configured boundaries, the algorithm steps down through acceptable values of each constraint in order of priority, from lowest to highest.

A satisfactory solution is defined as a set of ontology terms that meets a current minimum value for *overall_inclusion*, and in which every member meets the current minimum constraint for *r-score* and *specificity*. Upon each step of the algorithm, the subset of classes that meet the current individual criteria is identified, and if that set meets the current cutoff for *overall_inclusion*, this is returned as the solution. If not, the current constraint settings are stepped down in order of priority, with a new check for a satisfactory solution at each step. First, the cut-off for *r-score* is reduced stepwise to its lower limit, at which point the cut-off for *specificity* will be reduced by one step, and the cut-off for *r-score* will be reset. Once both *r-score* and *specificity* reach their lower limits, they are reset to their original values, and the *overall_inclusion* cut-off will be reduced by one step. This process is repeated until a satisfactory solution is found, checked every time a constraint is changed, or until the *overall_inclusion* cutoff reaches zero (a null response). The result of this process is a set of terms S_{good}, a subset of C, that both maximises the value of the objective function, and maximises values of the constraints according to the order of priority. The algorithm is given in Algorithm 1.

Data: Refer to Table 2.
Result: A set of ontology terms that explanatory of the cluster.
specificityCutoff = **top-ic**
r-scoreCutoff = **top-r-score**
totalInclusionCutoff = **top-total-inclusion**
while *True* **do**
  $S_p = \{$ x $\in U_j \mid$ *r-score*$(O_x, G_j) >=$ *rScoreCutoff* $\wedge$ *specificity*$(O_x) >=$ *specificityCutoff*$\}$
  $S_{pm} = \{$ x $\in S_p \mid$ *r-score*$(O_x, G_j) >=$ **max-r-score** $\wedge$ *specificity*$(O_x) >=$ **max-specificity** $\wedge$ *inclusion*$(O_x, G_j) >=$ **max-inclusion** $\wedge$ *exclusion*$(O_x, G_j) >=$ **max-exclusion**$\}$
  if *overall_inclusion*$(S_{pm}, G_j) <$ *totalInclusionCutoff* **then**
    if *rScoreCutoff* <= **bot-rScore then**
      rScoreCutoff = **top-rScore**
      if *specificityCutoff* > **bot-ic then**
        specificityCutoff = *specificityCutoff* - **step**
      else
        specificityCutoff = **top-ic**
        totalInclusionCutoff = *totalInclusionCutoff* - **step**
      end
    end
    rScoreCutoff = *rScoreCutoff* - **step**
  end
  return S_p **end**

Algorithm 1: Algorithm for identifying characterising ontology terms for a cluster, by stepping down through `r-score` and `specificity` thresholds.

The algorithm also supports maximal constraints, that are used to define the maximum score at which a candidate term can contribute to the *overall_inclusion* score. This is implemented in Algorithm 1 as the determination of the S_{pm} subset of the proposed solution S_p. By default, they are set to values that will not be triggered (shown in Table 2), but they can be configured to prevent the result set being dominated by a small number of highly explanatory classes, where a larger and more varied set is desired. For example, the appearance of the term `hypertension (HP:0008071)` at 0.95 inclusion for a group defined by a diagnosis of hypertension may not be very informative. Thus setting an upper limit to score values that can contribute to *overall_inclusion* will force the algorithm to continue to step down to find additional explanatory terms. Since these terms still characterise the group, however, they are still included in the explanatory set, and are therefore included in the output (and included in the final *overall_inclusion* value associated with the output). These upper parameters can be used to encourage the algorithm to seek larger, less monolithic explanatory sets.

While the upper and lower boundaries for cut-offs are configurable, reasonable defaults have been set based on our

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

observations using the algorithm. These can be defined for any of the term-wise scores: *r-score*, *inclusion*, *exclusion*, and *specificity*.

Table 2. Descriptions and default values for multivariable parameters for the stepdown algorithm.

Parameter	Default	Description
max-exclusion	1	Candidate terms with r-score values above this level won't contribute to calculation of overall inclusion, though they will still appear in explanatory sets.
max-inclusion	1	Candidate terms with r-score values above this level won't contribute to calculation of overall inclusion, though they will still appear in explanatory sets.
max-r-score	1	Candidate terms with r-score values above this level won't contribute to calculation of overall inclusion, though they will still appear in explanatory sets.
top-ic	0.7	The maximum stepdown value for IC.
bot-ic	0.4	The minimum stepdown value for IC.
top-r-score	0.8	The maximum stepdown value for r-score.
bot-r-score	0.1	The minimum stepdown value for r-score.
top-total-inclusion	1	Maximum stepdown value for overall inclusion.
step	0.05	Cutoffs will be reduced by this step on each iteration of the stepdown algorithm.

Discrimination and Significance

On top of the univariate and multivariate modes of producing explanatory sets of classes for groups, we have also implemented several additional methods by which the fitness of solutions can be evaluated. To identify the discriminatory power of a set of explanatory classes, we can define a function that will, for each combination of entity and group in the corpus, produce a score:

$$score(E_i, G_j, S_l) = \prod_{x=1}^{|S_l|}(1 \ + \ (exclusion(O_x, G_j) \ \cdot \ member(subclass(O_x), E_i))) \tag{11}$$

This score can either be calculated on the results of the multivariable stepdown approach, the full set of univariate results, or a manually specified list of term associations. In the case that we are using the univariate mode of execution only, S_l will be equivalent to U_j. Given a proposed solution for a group, consisting of a set of classes, a score is calculated for each entity in the corpus, defined as the product of one plus the exclusion score of every class in the solution set that the entity is annotated with.

This definition of a predictive score allows us to construct a model to classify group membership for entities, from which an Area Under the receiver operating Characteristic (AUC) score can be calculated. This provides a simple measure how well the set of explanatory variables, univariate or multivariate, was able to reclassify the set it was learned from. This can help to inform modifications to parameters or judge quality of the dataset in general.

The difference between these scores between the univariate and multivariate results can also provide insight on the quality of the modularisation performed by the stepdown algorithm. In the case that this value is small, there is a small amount of loss of total discrimination. Where values are large, there is a large decrease in performance when restricting scoring to the derived set of characteristic values, and this may be caused by a model that is poorly fit to the training data and does not reflect the full dataset.

The program also contains the possibility to apply this classifier to unseen data, calculating an AUC in the same way. This provides the ability to perform a test-set validation, or even external validation, to more properly judge the quality of the solution.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The second method of results evaluation is significance testing. In many instances, it is useful to access the type 1 error (false positive) rate to enable comparison with existing tools which provide p-values, and also to provide a conservative aid to interpreting inclusion and exclusion results. We implemented Monte Carlo methods to approximate the empirical p-value of the inclusion and exclusion statistics [38, 39]. Briefly, we sample with replacement a set of class associations for each entity of the same size as its original association set.

A test statistic (inclusion/exclusion) is generated for that set of patient profiles. This procedure is repeated 1,000 times to create a vector of test statistics. The statistics are ranked, and the empirical p-value is obtained as:

$$\hat{p} = \frac{r+1}{n+1} \tag{12}$$

where r is the number of test statistics ranked greater or equal to the observed value, and n is the total number of permutations. The full algorithm is shown in Supplementary Algorithm 1.

The number of permutations can be changed by the user, and we suggest at least 1,000 to ensure a (log-)normal distribution and sufficiently precise p-values possible (minimum of 0.002 in this case). Throughout the evaluations in this article, we use 2,000 permutations. Since we are creating a set of p-values for non-independent phenotypes, we suggest a Bonferroni threshold be applied to adjust the family-wise error rate:

$$\theta \leq \frac{\alpha_{fwe}}{n} \tag{13}$$

where αfwe is the desired error rate (0.05), n is the total number of phenotypes Klarigi generates p-values for (two for each term in the multivariable result set). Thus θ is the new threshold for a significant inclusion/exclusion score.

Output and Configuration

Klarigi reports its results in either the univariate or multivariate mode, in the form of a table of class associations with each group of interest, reported with the scores assigned to them. These can either be received in a LaTeXtable, a plain text table, or in Tab-Separated Values format.

The program is highly configurable, and we have previously discussed the parameters that are available to modify the operation and results of the program in both univariate (see Table 1) and multivariate modes of operation (see Table 2). In the former case, these parameters also affect the input of the multivariable analysis. In addition, there are additional parameters that control different modes of operation and calculation of scores. For example, there is a choice between which information content measure to use for the calculation of the *specificity* score. Descriptions of parameters are available in the documentation for the program. Klarigi can also be run in exclusive group loading (EGL) mode, which only loads the group of interest, and does not consider any others. In this case, the stepdown algorithm operates upon *inclusion*, and neither *exclusion* nor *r-score* are calculated. This can be useful for investigating group constituency without taking into account the exclusivity of terms.

Because of this configurability, and the dependence of results on that configurability, we anticipate a general workflow of using the application to consist of interactive modification of those parameters according to intermediate results, and desired outcomes. For example, if an empty set is returned upon analysis, the **min-ic** parameter could be reduced in an attempt to identify explanatory terms with a lower *specificity* than previously considered.

Results

We have described the development and implementation of Klarigi, and made the tool freely available at `https://github.com/reality/klarigi`. This repository includes the source code, pre-compiled binaries, documentation, and a tutorial in notebook format that walks through the functionality of the software, explaining the significance of different parameters and outputs.

To demonstrate and evaluate the use of Klarigi for exploration of biomedical datasets, we developed two use cases. Both describe clinical entities annotated using the Human Phenotype Ontology (HPO) [40]: in the first case text-derived phenotype profiles for a set of Medical Information Mart for Intensive Care III (MIMIC-III) admissions [41], and the second using a set of phenopackets [11], describing patients with rare diseases reported in literature [42].

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We use Klarigi to explore both of these datasets, comparing and contrasting those results with literature and enrichment analysis.

To further evaluate the results Klarigi produced for these use cases, we also used an enrichment method to identify over-represented terms. To do this, we used XGR [22] to identify over-represented phenotypes using both one-tailed binomial and fisher tests. P-values were adjusted with Bonferroni correction, propagating patient annotations to each term according to the True Path Rule. P-values were calculated with two background distributions: once using traits annotated to each patient and their frequencies as background, and additionally using only those annotated to the parent class of each entity tested; the maximum of each p-value was selected to help correct for the interdependence of testing. In the case of Klarigi, we also tested the normality of the permutation distribution for inclusion and exclusion scores by visually examining the distributions using histograms and QQ-plots. An R script containing these visualisations is provided in experiment repository.

The code used to obtain our results, as well as to perform the quantitative evaluations and create the tables for the paper, are available online at `https://github.com/reality/klarigi_paper`.

Use Case: Pulmonary Embolism

Pulmonary embolism, a condition associated with considerable mortality rates, presents in ways that render it difficult to diagnose when associated with other comorbidities, such as COPD, and typically shares symptoms with other more common conditions, such as pneumonia and acute bronchitis [43]. The critical time dependence of treatment on diagnosis makes it important to identify combinations of discriminating symptoms as rapidly as possible [44].

To demonstrate Klarigi's functionality, and to gain insight into the phenotypic presentations associated with pulmonary embolism and pneumonia, we created and evaluated text-derived phenotype profiles describing critical care admissions with the two diagnoses. To do this, we identified relevant admissions from MIMIC-III, and analysed their associated clinical narrative. MIMIC-III is a large freely accessible dataset concerning nearly 60,000 visits to the emergency department at the Beth Israel Deaconess hospital [41].

We used the ICD-9 codes `486` and `41519` to identify 5,437 MIMIC-III admissions with either pneumonia or pulmonary embolism as an associated diagnosis (provided in the coded diagnosis list produced by a coding expert): 4,853 with pneumonia and 912 with pulmonary embolism. We further restricted these to admissions with pneumonia or pulmonary embolism given as the primary diagnosis, and those which did not have the other diagnosis in any position. This led to a final set of 991 admissions, made up of 699 primarily coded with pneumonia, and 292 with pulmonary embolism. We employed the 2021-08-02 release of the Human Phenotype Ontology, available at `http://purl.obolibrary.org/obo/hp/releases/2021-08-02/hp.owl`.

To create phenotype profiles for these admissions, we used the Komenti semantic text-mining framework [45], which implements Stanford CoreNLP [46]. For every considered admission, we collected the text from the corresponding discharge note. We created a lemmatised vocabulary of all labels and synonyms in the Human Phenotype Ontology (HPO) [40], containing 50,265 labels for 16,019 unique classes. We used these to identify ontology terms mentioned in the discharge notes. We then removed all associations that were identified by Komenti as negated or uncertain. We also removed classes equivalent to or subclasses of `pneumonia (HP:0002090)` and `pulmonary edema (HP:0100598)` from the set of annotations. To facilitate a holdout validation, we then reserved a randomly sampled 20% of the annotated admissions. In the training set there are a total of 785 records, 552 pneumonia and 233 pulmonary embolism, while in the test set there were 206 records, 147 pulmonary embolism and 59 pneumonia.

We first employed Klarigi in "exclusive group loading" (EGL) mode separately on both groups. As described in the implementation section, this evaluates the selected group alone, without considering its relationship to any other groups in the dataset. As such, only the inclusion and IC scores are used, and the step-down algorithm uses inclusion rather than r-score as its optimisation parameter. The results of this analysis are shown in Supplementary Table 1. For both conditions, large values of the overall inclusion score show that the given explanatory sets describe the cohort, although in the case of pulmonary embolism this set is much larger, at 41 explanatory terms, relative to the 10 identified for pneumonia.

To gain insight on discriminatory differences between these two groups, Klarigi was used in its default mode: with the opposing group as context, providing the additional *exclusion* and *r-score* measures, using the latter for the stepdown algorithm this time. The full set of univariate results for each disease are given in Supplementary Tables 2 and 3, while the multivariable results derived from the stepdown algorithm performed on the full set of univariate

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

results, and additionally permutation tested, are shown in Table 3. We modified the `min-r-score` and `min-exclusion` hyperparameters manually to optimise for the reclassification AUC on the training set, whose final values were 0.754 and 0.838 for pulmonary embolism and pneumonia respectively. Permutation testing was performed after hyper-parameter optimisation, and so was only run once. Meanwhile, the results of the enrichment analysis using both the Fisher and binomial tests are shown in Supplementary Table 4.

Table 3. Multivariable Klarigi results for pneumonia and pulmonary embolism, with each others' cases considered as control. Train AUC was 0.754 for pulmonary embolism and Train AUC was 0.838 for pneumonia.

pulmonary embolism (233 members)	**r-score**	**Inclusion**	**Exclusion**	**IC**
Abnormal thrombosis (HP:0001977)	0.37	0.41 (p=0.001)	0.35 (p=0.001)	0.75
Venous thrombosis (HP:0004936)	0.35	0.37 (p=0.001)	0.34 (p=0.001)	0.83
Deep venous thrombosis (HP:0002625)	0.33	0.33 (p=0.001)	0.34 (p=0.002)	1.0
Hypercoagulability (HP:0100724)	0.19	0.11 (p=0.001)	0.7 (p=0.001)	1.0
Lower limb pain (HP:0012514)	0.18	0.14 (p=0.001)	0.26 (p=0.004)	0.89
Limb pain (HP:0009763)	0.18	0.14 (p=0.001)	0.25 (p=0.006)	0.84
Pleuritic chest pain (HP:0033771)	0.17	0.15 (p=0.001)	0.22 (p=0.01)	1.0
Increased body weight (HP:0004324)	0.17	0.15 (p=0.001)	0.19 (p=0.024)	0.84
Sinus tachycardia (HP:0011703)	0.17	0.25 (p=0.001)	0.13 (p=0.069)	1.0
Obesity (HP:0001513)	0.15	0.13 (p=0.001)	0.17 (p=0.042)	0.87
Abnormality of skin adnexa physiology (HP:0025276)	0.12	0.12 (p=0.001)	0.13 (p=0.058)	0.83
Hyperhidrosis (HP:0000975)	0.12	0.12 (p=0.001)	0.13 (p=0.058)	0.95
Hypohidrosis or hyperhidrosis (HP:0007550)	0.12	0.12 (p=0.001)	0.13 (p=0.058)	0.92
Abnormal exteroceptive sensation (HP:0033747)	0.12	0.09 (p=0.001)	0.16 (p=0.043)	0.8
Somatic sensory dysfunction (HP:0003474)	0.12	0.09 (p=0.001)	0.16 (p=0.043)	0.75
Syncope (HP:0001279)	0.12	0.09 (p=0.001)	0.16 (p=0.036)	0.89
Thromboembolism (HP:0001907)	0.12	0.07 (p=0.001)	0.4 (p=0.001)	0.89
Hiatus hernia (HP:0002036)	0.11	0.08 (p=0.001)	0.18 (p=0.028)	1.0
Hypoesthesia (HP:0033748)	0.11	0.08 (p=0.001)	0.16 (p=0.043)	1.0
Overall	-	*79.4*	-	-
pneumonia (552 members)	**r-score**	**Inclusion**	**Exclusion**	**IC**
Renal insufficiency (HP:0000083)	0.22	0.39 (p=0.001)	0.15 (p=0.006)	0.84
Airway obstruction (HP:0006536)	0.2	0.34 (p=0.001)	0.14 (p=0.027)	0.87
Atrial fibrillation (HP:0005110)	0.2	0.27 (p=0.001)	0.15 (p=0.074)	0.95
Respiratory insufficiency (HP:0002093)	0.2	0.24 (p=0.001)	0.17 (p=0.012)	0.81
Chronic pulmonary obstruction (HP:0006510)	0.2	0.33 (p=0.001)	0.14 (p=0.092)	1.0
Respiratory distress (HP:0002098)	0.2	0.37 (p=0.001)	0.13 (p=0.125)	0.92
Congestive heart failure (HP:0001635)	0.2	0.32 (p=0.001)	0.14 (p=0.096)	0.95
Aspiration (HP:0002835)	0.19	0.29 (p=0.001)	0.15 (p=0.095)	1.0
Respiratory failure (HP:0002878)	0.19	0.22 (p=0.001)	0.17 (p=0.049)	1.0
Productive cough (HP:0031245)	0.19	0.17 (p=0.001)	0.22 (p=0.018)	1.0
Chronic kidney disease (HP:0012622)	0.18	0.17 (p=0.001)	0.21 (p=0.002)	0.87
Atrial arrhythmia (HP:0001692)	0.18	0.29 (p=0.001)	0.13 (p=0.05)	0.83
Rhonchi (HP:0030831)	0.18	0.27 (p=0.001)	0.13 (p=0.128)	1.0
Stage 5 chronic kidney disease (HP:0003774)	0.17	0.13 (p=0.001)	0.24 (p=0.006)	1.0
Confusion (HP:0001289)	0.17	0.18 (p=0.001)	0.15 (p=0.024)	1.0
Sepsis (HP:0100806)	0.16	0.17 (p=0.001)	0.15 (p=0.09)	1.0
Respiratory tract infection (HP:0011947)	0.15	0.13 (p=0.001)	0.18 (p=0.055)	0.73
Abnormality of the pharynx (HP:0000600)	0.15	0.12 (p=0.001)	0.19 (p=0.038)	0.74
Pharyngalgia (HP:0033050)	0.14	0.11 (p=0.001)	0.18 (p=0.046)	1.0
Abnormal pharynx physiology (HP:0033152)	0.14	0.11 (p=0.001)	0.18 (p=0.046)	1.0
Nonproductive cough (HP:0031246)	0.13	0.09 (p=0.001)	0.22 (p=0.017)	1.0
Delirium (HP:0031258)	0.12	0.1 (p=0.001)	0.17 (p=0.067)	1.0
Abnormal bronchus morphology (HP:0025426)	0.12	0.1 (p=0.001)	0.16 (p=0.001)	0.72
Malaise (HP:0033834)	0.12	0.08 (p=0.001)	0.23 (p=0.012)	1.0
Hepatitis (HP:0012115)	0.11	0.09 (p=0.001)	0.14 (p=0.121)	0.81
Abnormality of the abdominal wall (HP:0004298)	0.1	0.08 (p=0.001)	0.16 (p=0.001)	0.73
Overall	-	*95.47*	-	-

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To evaluate how characteristic and discriminatory the explanatory sets were of their respective diseases, we used the 20% holdout set to define a classification task. The results of this analysis are listed in Table 4. In this analysis, we used the sets of characteristic terms identified by Klarigi in univariate mode, Klarigi in multivariate mode, and enrichment. Since the binomial enrichment results were a subset of the Fisher results, we combine these into one 'enrichment' set of explanatory terms. We then constructed classifiers for pulmonary embolism and pneumonia in the test set with scores for each entity identified using Klarigi. To provide additional measures of evaluation, as well as to evaluate the utility of the exclusion score in measuring term discrimination, we also compared the Klarigi classification method with semantic similarity approaches. To do this, we also built classifiers scored using the groupwise similarity between each phenotype profile in the test set, and the explanatory set for each method. We performed this task for both the best match average and average groupwise strategies, with both cases using the Resnik pairwise method and Resnik information content measure (using the annotations in the training set for the probability distribution). For all classifiers, we calculated an AUC using the scores for each admission in the test set, and their actual diagnosis group for the label.

Table 4. Test set classification results for pneumonia and pulmonary embolism, using Klarigi's classifier (see Equation 11), and semantic similarity using the Resnik IC and Resnik pairwise method, with results given for both best match average and average groupwise strategies. OI is the overall inclusion score, defined in Equation 10, of the solution with respect to the test set. The greatest score for each category is emboldened.

Group	Method	AUC (Klarigi)	AUC (SS-BMA)	AUC (SS-AVG)	OI
Pneumonia	Enrichment	0.819	0.66	0.694	0.939
	Klarigi (u)	**0.833**	**0.668**	**0.705**	**0.952**
	Klarigi (mv)	0.83	0.665	**0.705**	**0.952**
Pulmonary embolism	Enrichment	0.603	0.587	0.556	0.102
	klarigi (u)	0.672	0.661	0.67	**0.661**
	Klarigi (mv)	**0.677**	**0.663**	**0.676**	**0.661**

Use Case: Phenopackets

Phenopackets are a standardised format for the representation of phenotypic descriptions of patients, linked to biomedical ontologies [11]. We developed Klarigi to natively support the phenopackets format, converting it internally into the required data model. To further evaluate Klarigi, we examined a dataset of 384 rare disease patients, whose phenotypes were originally reported in literature, and have subsequently had their phenotypes transcribed into the phenopacket format [42]. The dataset describes a range of rare diseases, and in our use case, we examine the disease with the most patient descriptions, `Hypotonia, infantile, with psychomotor retardation and characteristic facies 3 (OMIM:616900)` (IHPRF3), with 19 patients. In the dataset, the patient descriptions are derived from 4 separate publications, listed in Table 5 were taken from these citations. The 2018-03-08 release of HPO was used to annotate these patients, and this was the version we used for the analysis.

Table 5. The publications that report the 19 patients with IHPRF3 that are phenotyped in the phenopackets dataset.

Short Name	Full Title	Patients
Bhoj 2016 [47]	Mutations in TBCK, Encoding TBC1-Domain-Containing Kinase, Lead to a Recognizable Syndrome of Intellectual Disability and Hypotonia	10
Chong 2016 [48]	Recessive Inactivating Mutations in TBCK, Encoding a Rab GTPase-Activating Protein, Cause Severe Infantile Syndromic Encephalopathy	5
Guerreiro 2016 [49]	Mutation of TBCK causes a rare recessive developmental disorder	2
Zapata-Aldana 2019 [50]	Further delineation of TBCK - Infantile hypotonia with psychomotor retardation and characteristic facies type 3	2
Total	-	19

Supplementary Table 5 shows the full set of 61 associated classes that constitute the univariate results.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Meanwhile, Table 6 shows the multivariable results, whose overall inclusion describes complete coverage of the dataset with at least one of these classes, expressed in a much smaller explanatory set of 33 classes. Enrichment analysis was also performed using the Fisher and binomial tests, and the significant results of this analysis are shown in Supplementary Table 6.

Table 6. Multivariable results for IHPRF3 in Klarigi. Reclassification AUC was 0.967.

OMIM:616900 (19 members)	**r-score**	**Inclusion**	**Exclusion**	**IC**
Severe muscular hypotonia (HP:0006829)	0.8	0.68 (p=0.001)	0.95 (p=0.01)	1.0
Developmental regression (HP:0002376)	0.63	0.58 (p=0.001)	0.68 (p=0.009)	1.0
Severe global developmental delay (HP:0011344)	0.5	0.42 (p=0.001)	0.62 (p=0.01)	1.0
Abnormality of upper lip vermillion (HP:0011339)	0.46	0.47 (p=0.001)	0.45 (p=0.007)	0.83
Respiratory insufficiency (HP:0002093)	0.42	0.42 (p=0.001)	0.42 (p=0.041)	0.81
Reduced tendon reflexes (HP:0001315)	0.42	0.53 (p=0.001)	0.35 (p=0.019)	0.82
Small basal ganglia (HP:0012697)	0.41	0.26 (p=0.001)	0.95 (p=0.009)	1.0
Exaggerated cupid's bow (HP:0002263)	0.41	0.26 (p=0.001)	0.95 (p=0.008)	1.0
Macroglossia (HP:0000158)	0.39	0.26 (p=0.001)	0.78 (p=0.008)	0.95
Areflexia (HP:0001284)	0.39	0.32 (p=0.001)	0.5 (p=0.013)	0.89
Profound global developmental delay (HP:0012736)	0.36	0.26 (p=0.001)	0.58 (p=0.012)	1.0
Prominent nasal bridge (HP:0000426)	0.36	0.26 (p=0.001)	0.58 (p=0.008)	1.0
Skeletal muscle hypertrophy (HP:0003712)	0.35	0.26 (p=0.001)	0.51 (p=0.011)	0.78
Sloping forehead (HP:0000340)	0.34	0.21 (p=0.001)	0.95 (p=0.008)	1.0
Aplasia/Hypoplasia of the cerebellar vermis (HP:0006817)	0.33	0.26 (p=0.001)	0.45 (p=0.038)	0.87
Highly arched eyebrow (HP:0002553)	0.33	0.26 (p=0.001)	0.45 (p=0.047)	1.0
Cerebellar vermis hypoplasia (HP:0001320)	0.33	0.26 (p=0.001)	0.45 (p=0.038)	0.95
Tented upper lip vermilion (HP:0010804)	0.31	0.21 (p=0.001)	0.62 (p=0.009)	1.0
Coarse facial features (HP:0000280)	0.29	0.26 (p=0.001)	0.34 (p=0.038)	1.0
Small forehead (HP:0000350)	0.29	0.26 (p=0.001)	0.34 (p=0.044)	1.0
Narrow forehead (HP:0000341)	0.29	0.26 (p=0.001)	0.34 (p=0.044)	1.0
Aplasia/Hypoplasia of the corpus callosum (HP:0007370)	0.29	0.37 (p=0.001)	0.24 (p=0.083)	0.95
Aplasia/Hypoplasia of the cerebral white matter (HP:0012429)	0.29	0.37 (p=0.001)	0.24 (p=0.083)	0.95
Hypoplasia of the corpus callosum (HP:0002079)	0.29	0.32 (p=0.001)	0.27 (p=0.083)	1.0
Cerebral white matter hypoplasia (HP:0012430)	0.29	0.32 (p=0.001)	0.27 (p=0.083)	1.0
Brachycephaly (HP:0000248)	0.29	0.21 (p=0.001)	0.45 (p=0.043)	0.95
Absent speech (HP:0001344)	0.28	0.37 (p=0.001)	0.23 (p=0.08)	1.0
Abnormality of the cerebellar vermis (HP:0002334)	0.27	0.26 (p=0.001)	0.28 (p=0.035)	0.8
Diffuse cerebellar atrophy (HP:0100275)	0.27	0.16 (p=0.001)	0.95 (p=0.008)	1.0
Global brain atrophy (HP:0002283)	0.27	0.16 (p=0.001)	0.95 (p=0.008)	1.0
Global developmental delay (HP:0001263)	0.26	0.95 (p=0.001)	0.15 (p=0.057)	0.89
Open mouth (HP:0000194)	0.26	0.16 (p=0.002)	0.7 (p=0.008)	1.0
Diffuse cerebral atrophy (HP:0002506)	0.26	0.16 (p=0.005)	0.7 (p=0.001)	1.0
Visual impairment (HP:0000505)	0.25	0.21 (p=0.001)	0.31 (p=0.049)	0.87
Overall	-	*100.0*	-	-

To provide a quantitative evaluation for this much smaller set of patients, we compared them with the disease annotations defined by the HPO database. HPO database annotations were developed through a combination of expert curation and text mining of clinical descriptions provided in OMIM [51]. These relationships are produced by expert curators, and are derived from literature, clinical descriptions, or experimental evidence. There are 46 such annotations, listed in Supplementary Table 4. To measure how characteristic the result sets are we calculated semantic similarity between each one and the set of HPO database annotations. We used the Resnik method of pairwise similarity, with the Zhou method of information content, with the best match average groupwise strategy. The results of the Fisher and Binomial tests are here treated separately, since neither forms a subset of the other.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 7 lists the semantic similarity value for each comparison.

Table 7. Resnik+BMA+Zhou semantic similarity between the the explanatory set for each method and the definitive set of IHPRF3 HPO annotations defined by the HPO annotations database.

Method	Similarity
Fisher	0.560
Binomial	0.607
Klarigi (u)	0.703
Klarigi (mv)	**0.72**

We can also use Klarigi to evaluate alternative ways of grouping the data, or subsets of the data. In this use case, we explored 19 IHPRF3 patient data across different publications, shown in Table 5. In Supplementary Table 7 we show multivariable Klarigi results for IHPRF3 patients grouped by the publication in which they are reported.

Discussion

We have described the design and implementation of Klarigi, and its application to two use cases, one to the comparison of two phenotypically similar diseases from text-derived phenotypes, and another to rare disease patients from expert annotations created from literature. In this section, we will interpret and discuss the results, and provide more discussion of Klarigi in context, including limitations and future work.

Interpretation of Results

Pulmonary Embolism

The large values of overall inclusion shown for the EGL results in Supplementary Table 1 show that the explanatory sets had a large coverage, meeting the overall inclusion threshold before being reduced, per the stepdown algorithm. The much greater number of explanatory terms for pulmonary embolism implies a more complex and varied phenotypic presentation for patients with this disease, since a greater number of explanatory terms with a lower individual threshold for inclusion had to be included to reach the overall inclusion threshold. This difference illustrates how the stepdown algorithm works to provide explanations that suit the data being described, beginning with thresholds that admit parsimonious solutions with large values for individual scores, lowering those thresholds as necessary for more complex, noisy, or difficult to characterise entity groups.

The EGL results also evidence the similarity in presentation of the two conditions, with the 9 terms listed in the pneumonia set forming an explicit subset of those appearing in the pulmonary embolism set: every single explanatory term for pneumonia is also one for pulmonary embolism. The similarities are explicated further by the similarities in the representation of those terms, evidenced by the inclusion scores. The mean difference in inclusion score between PE and pneumonia for the 9 explanatory terms for pneumonia is only 0.13, though some individual terms, such as `cough (HP:0012735)` and `abnormal breath sound (HP:0030829)`, have greater differences in inclusion score, implying that they may be candidates for discriminatory terms. In fact, the enrichment results include `cough (HP:0012735)`, being the only significant result produced by the binomial method.

The complication of this dataset, illustrated by the EGL results, demonstrates the challenge of the task in finding discriminatory characteristic sets for these diseases. The full results of the multivariate analysis performed with consideration of the other group, shown in Table 3, works to create such a solution. Indeed, none of the other highly redundant 9 terms in the EGL results appear in either set of results, however a more specific, and in this case discriminatory, subclass `productive cough (HP:0031245)` appears in the set of explanations for pneumonia.

The reclassification AUCs for both reveal good performance, indicating that both sets are reasonably characteristic. This was lower for pulmonary embolism, which mirrors the lower overall inclusion score, and is indicative of the increased complexity of this disease and its presentation in comparison to pneumonia.

The discriminatory power of the sets is further explored by the classification analysis, the results of which are given in Table 4. Reduction in performance for prediction of pneumonia with the Klarigi classifier was minimal, indicating that the set of explanatory terms generalised well to the phenotypic distribution of those patients in the test set. This

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reduction was much greater for pulmonary embolism, which tracks the lower overall quality of this explanatory set evident on the training set, although the classifier still achieved reasonable performance with a strong signal.

For all classification methods and for both groups, models derived from the sets of explanatory terms generated by Klarigi outperformed those from enrichment. This is very stark in the case of pulmonary embolism, where performance of the enrichment on the test set approached random, with poor generalisation also evidenced by the very low overall inclusion value.

Another insight from the classification analysis on the test set, is that the Klarigi classifier consistently performed substantially better than both semantic similarity approaches. This helps to provide evidence that the exclusion score, which makes up the Klarigi classification model, is a good measure of term discrimination, that generalises to use for quantifying discriminatory sets.

Phenopackets

The phenopackets dataset is much smaller than the MIMIC dataset, describing far fewer patients. The annotations of patients were also derived by an expert from literature reports, instead of being mined from text. The task also differs, in that we are evaluating IHPRF3 patients against a large background distribution of patients with other rare genetic diseases, instead of comparing their phenotype profiles head-to-head with those of a single chosen, and phenotypically similar disease. This difference is evidenced by the increased clarity of the results, with large individual values for inclusion and exclusion, overall inclusion, and reclassification AUC. Klarigi was also able to identify a much smaller subset of the univariate sets during the multivariate stage, and this provides an example of the identification of sparse representations of entity groups, in the case that the representation of terms across the group is sufficiently homogenous.

The small number of disease cases also informed the decision not to perform a train-test validation for this use case. Instead, explanatory sets were compared with semantic similarity to the HPO database annotations for the disease. In this evaluation, Klarigi's result sets were shown to be more similar to the definitional annotations than were both enrichment approaches.

We can also use this dataset to identify relationships that are not represented in the structured description of the disease. Several classes identified by the multivariate results, described in Table 6, are not found in the HPO database annotations. For instance, none of the phenotypes describing aplasia/hypoplasia are included in the OMIM database, such as `Aplasia/Hypoplasia of the corpus callosum (HP:0007370)`, which affects 37% of described patients with IHPRF3. As such, the findings produced by Klarigi could constitute new disease associations for the disease, not currently expressed in existing scientific databases describing it. In addition, of the HPO database annotations that are shared by Klarigi's representation, we provide additional information about patient representation, information content, and how exclusive these phenotypes are for patients across a wide set of rare diseases, all information that can be used to enrich secondary use of these phenotype associations.

Supplementary Table 8 shows the multivariable Klarigi results for IHPRF3 when stratified by the publication they appear in. This analysis allows us to identify features that may be more specific to particular publications. For example, `intellectual disability (HP:0001249)` is over-represented in the patients described by Bhoj 2016, and both listed phenotypes for Guerreiro 2016 are highly exclusive and inclusive. The terms appearing in these sets are highly discriminatory, and thus in fact may imply that they are not overall representative of the disease, and instead a feature or focus of the investigation described by the source article. Interestingly, we can see from the metadata associated with the HPO database annotations, that some associations were directly evidenced by the OMIM clinical description, which in turn makes use of publications not represented in the phenopackets dataset, in particular Alazami 2015 [52].

Comparison to Enrichment Analysis

The results of Klarigi cannot easily be directly compared to enrichment analysis, since enrichment analysis methods do not determine multivariable sets of explanatory variables, and Klarigi produces its results through class scoring rather than a significance cut-off. Indeed, the purposes of enrichment and Klarigi are different. In the former case we are aiming to identify significant individual assocations between terms and groups, while in the latter we are attempting to identify sets of characteristic and discriminatory terms. Thus, while our quantitative evaluations show that Klarigi out-performs enrichement in all cases, this primarily indicates the different problem that Klarigi solves:

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

identifying characteristic and discriminatory sets of terms. Klarigi's univariate mode is more comparable to enrichment, in that it provides a set of pairwise relationships between terms and the group of interest, though Klarigi provides more information about those relationships.

While enrichment analysis correctly identifies over-represented relationships, it provides little information about how characteristic the given term is with respect to the considered group. For instance, both Fisher and Binomial test, produce `abnormal thrombosis (HP:0001977)` as the term with the largest z-score and odds ratio in the small set of significant terms identified for pulmonary embolism. This association is not incorrect, though while this term is highly over-represented, and also appears in the Klarigi result sets, less than half of the admissions in the pulmonary embolism group are associated with it - indicated by the *inclusion* score of 0.41. In cases like this, large values for zscore and odds ratio may be confusing or misleading to interpreters, and does not provide enough information to those attempting to gain a holistic understanding of the group. Klarigi captures this combination of facts in its scores for `abnormal thrombosis (HP:0001977)`, with a large score for exclusion, small score for inclusion, and balanced r-score metric. Moreover, in the Klarigi results, it appears in the context of other terms with which it does form a more complete phenotypic characterisation of the group, as evidenced by the performance of the classifier on the test set.

Since we are not selecting multivariable terms based on significance testing, this gives us the opportunity to manually and reactively modify hyper-parameters to identify better explanatory modules for the group. For example, if initial results are too general, or if the result set is dominated by a single class, the *max-inclusion* parameter can be modified. In the case of the pulmonary embolism use case, we used the *min-exclusion* and *min-r-score* parameters to improve discrimination between the groups.

Limitations and Future Work

While one feature of Klarigi is the ability to interactively modify hyperparameters to identify desirable result sets, these parameters also introduce complexity to the process, and understanding the interplay of the parameters and their effect on the results may require manual work. Grid search could also be employed to automatically identify good settings for parameters, although a challenge here lies in identifying a good optimisation heuristic, as the current `overall_inclusion` may be equivalent across result sets of different quality, with respect to a desired outcome for the human operator, depending on their research goals - for example, five terms that describe 100% of the group may be preferable to one term that also describes 100% of the group. Such a search through possible overall parameters could also add significantly to the execution time. However, overcoming this problem would also help to avoid the problem of providing parameter defaults that perform well across different ontologies, datasets, and algorithmic parameters, and reduce the amount of human input required to obtain quality results. Parameter optimisation could also potentially introduce problems with multiple testing. We controlled for this, however, by choosing not to employ permutation testing until the final set of desired results is received - Klarigi currently requires an additional parameter to be passed for significance testing to be performed.

It is also possible that different scores could be introduced, or different definitions of existing scores could be explored. For example, the option to re-weight the r-score, in a similar manner to differential weight applications to precision and recall in an f-score.

The information Klarigi produces about unique characterisation of rare diseases may also benefit development of disease-phenotype associations, which are used in a wide range of analysis approaches, such as rare disease variant prioritisation or diagnosis. This contribution could consist simply in the identification of additional phenotypes through analysis of either larger datasets similar to the phenopackets dataset, or text-mined data from clinical encounters or literature. Moreover, we showed in Supplementary Table 8 that we were able to identify disease phenotypes that were more exclusively associated with single publications. Taking note of this information could support development of high quality disease-phenotype relationships. For example, particular phenotypes that are very exclusive to one or few publications may be less preferred as candidates for associations, or associations could be provided with a measure of inter-literature agreement. Such a dataset could be produced using a larger set of phenopackets data, or through mining literature and clinical narrative for phenotype profiles. Klarigi parameters could be configured to identify the necessary level of association desired to constitute a disease-phenotype link.

The measures described in our experiment could also be combined with enrichment analysis approaches. Enrichment analysis use a measure of *enrichment score*, which could comprise any of the r-score, inclusion, or exclusion scores given herein, then using the usual method of testing for significant effects. Indeed, the use of

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

significance testing upon the final set of multivariable results can be considered as a kind of enrichment analysis. Other potential expansions of functionality could consist in extending the tool to facilitate a wider range of information content measures, involving multi-faceted semantic similarity, or identification of sub-groups through clustering.

Admissions sampling for the MIMIC dataset did not consider whether multiple admissions were attributed to the same patient, and this could potentially be a small source of bias. However, we do not believe it affects the conclusions of this study, since it would not favour a particular algorithm. In the same way, different publications described by the phenopackets dataset could potentially describe the same patients, although there is no way to confirm this programmatically.

We did not compare the approach with another multivariable enrichment method. This is because existing methods for multivariable enrichment are either limited to the Gene Ontology, or require a complicated setup process. The possibility for Klarigi to be easily applied to any ontology and ontology-annotated dataset is a major benefit of the approach. We addressed this by making these methodological differences clear in our problem description, comparison of the results, and discussion.

We anticipate that Klarigi can be used to perform improved semantic descriptive analysis of biomedical entities. This is an area of research that is of recent interest, especially with advancements in text mining for the automated production of phenotypic profiles. For example, one study performed a statistical analysis of text-mined phenotypic profiles for 4,095 individuals with Down Syndrome [53], descriptively reporting on phenotypes and their frequency of appearance. This analysis, however, only used a measure of patient frequency to stratify phenotypes. The use of Klarigi in this case would have introduced measures of representation for considered phenotypes, for example overall coverage of unique individuals, and the utility to easily switch around groupings to explore the different aspects of the data.

Conclusions

Klarigi provides a new way to solve the task of semantic explanation, and thus to explore and interpret biomedical datasets linked to ontologies. We have demonstrated that it can be used to gain insight into relationships between biomedical entities with clinical relevance.

Acknowledgements

Funding

GVG and LTS acknowledge support from the NIHR Birmingham ECMC, NIHR Birmingham SRMRC, Nanocommons H2020-EU (731032), MAESTRIA (Grant agreement ID 965286), HYPERMARKER (Grant agreement ID 101095480), and the MRC Heath Data Research UK (HDRUK/CFC/01), an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council or the Department of Health. RH and GVG were supported by funding from King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3790-01-01. AK was additionally supported by the Medical Research Council (MR/S003991/1).

Abbreviations

HPO: Human Phenotype Ontology, IHPRF3: Hypotonia, infantile, with psychomotor retardation and characteristic facies 3, OMIM: Online Mendelian Inheritance In Man, OWL: Web Ontology Language, GSEA: Gene Set Enrichment Analysis, EGL: Exclusive Group Loading, GO: Gene Ontology, AUC: Area Under receiver operating Characteristic, IC: Information Content, MIMIC-III: Medical Information Mart for Intensive Care III.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ethics approval and consent to participate

This work makes use of the MIMIC-III dataset, which was approved for construction, de-identification, and sharing by the BIDMC and MIT institutional review boards (IRBs). Further details on MIMIC-III ethics are available from its original publication (DOI:10.1038/sdata.2016.35). Work was undertaken in accordance with the MIMIC-III guidelines.

Competing interests

John Williams is an employee of Eisai, Inc. Eisai, Inc had no role in funding or design of this study.

Authors' contributions

LTS conceived of the study, designed and implemented the software, and created the initial manuscript. **JAW** contributed to design of the software, experimental design, and the manuscript, particularly surrounding the statistical analysis and significance testing. **AK** contributed to the manuscript and experimental design. **SR** tested multiple prototypical versions of the software and provided feedback that contributed to the final design. **PNS** performed clinical and biological interpretation of results, contributed to the manuscript and experimental design. **HF and SB** contributed to project supervision and experimental design. **RH and GVG** supervised the project, contributed to experimental design, and contributed to the manuscript. All authors approved the manuscript for submission.

References

1. Hoehndorf R, Schofield PN, Gkoutos GV. PhenomeNET: A Whole-Phenome Approach to Disease Gene Discovery. Nucleic Acids Research. 2011 Oct;39(18):e119–e119.

2. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. The American Journal of Human Genetics. 2009 Oct;85(4):457–464.

3. Slater LT, Karwath A, Williams JA, Russell S, Makepeace S, Carberry A, et al. Towards Similarity-Based Differential Diagnostics for Common Diseases. Computers in Biology and Medicine. 2021 Jun;133:104360.

4. Veatch OJ, Veenstra-Vanderweele J, Potter M, Pericak-Vance MA, Haines JL. Genetically Meaningful Phenotypic Subgroups in Autism Spectrum Disorders. Genes, Brain, and Behavior. 2014 Mar;13(3):276–285.

5. Kulmanov M, Smaili FZ, Gao X, Hoehndorf R. Machine Learning with Biomedical Ontologies. bioRxiv. 2020 May;p. 2020.05.07.082164.

6. Gentleman R, Morgan M, Huber W. Gene set enrichment analysis. In: Bioconductor Case Studies. Springer; 2008. p. 193–205.

7. Deng Y, Gao L, Wang B, Guo X. HPOSim: An R Package for Phenotypic Similarity Measure and Enrichment Analysis Based on the Human Phenotype Ontology. PLOS ONE. 2015 Feb;10(2):e0115692.

8. Verspoor K. The Gene Ontology as a Source of Lexical Semantic Knowledge for a Biological Natural Language Processing Application.; 2003. .

9. Kazakov Y, Krötzsch M, Simančík F. The Incredible ELK. Journal of Automated Reasoning. 2014 Jun;53(1):1–61.

10. Slater LT, Williams JA, Karwath A, Fanning H, Ball S, Schofield P, et al. Multi-Faceted Semantic Clustering With Text-derived Phenotypes. medRxiv. 2021 May;p. 2021.05.26.21257830.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

11. Jacobsen JOB, Baudis M, Baynam GS, Beckmann JS, Beltran S, Callahan TJ, et al. The GA4GH Phenopacket Schema: A Computable Representation of Clinical Data for Precision Medicine. medRxiv. 2021 Nov;p. 2021.11.27.21266944.

12. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast Gene Set Enrichment Analysis; 2021.

13. Morgan M, Falcon S, Gentleman R. GSEABase: Gene Set Enrichment Data Structures and Methods; 2021. Bioconductor version: Release (3.14).

14. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: Tool for the Unification of Biology. Nature Genetics. 2000 May;25(1):25–29.

15. Yu G, Wang LG, Yan GR, He QY. DOSE: An R/Bioconductor Package for Disease Ontology Semantic and Enrichment Analysis. Bioinformatics. 2015 Feb;31(4):608–609.

16. Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, et al. Disease Ontology: A Backbone for Disease Semantic Integration. Nucleic Acids Research. 2012 Jan;40(D1):D940–D946.

17. Yu G, He QY. ReactomePA: An R/Bioconductor Package for Reactome Pathway Analysis and Visualization. Molecular BioSystems. 2016 Jan;12(2):477–479.

18. Weng MP, Liao BY. MamPhEA: A Web Tool for Mammalian Phenotype Enrichment Analysis. Genome Biology. 2010 Oct;11(1):P27.

19. Smith CL, Goldsmith CAW, Eppig JT. The Mammalian Phenotype Ontology as a Tool for Annotating, Analyzing and Comparing Phenotypic Information. Genome Biology. 2004 Dec;6(1):R7.

20. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, et al. DAVID Bioinformatics Resources: Expanded Annotation Database and Novel Algorithms to Better Extract Biology from Large Gene Lists. Nucleic Acids Research. 2007 Jul;35(suppl_2):W169–W175.

21. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. G:Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update). Nucleic Acids Research. 2019 Jul;47(W1):W191–W198.

22. Fang H, Knezevic B, Burnham KL, Knight JC. XGR Software for Enhanced Interpretation of Genomic Summary Data, Illustrated by Application to Immunological Traits. Genome Medicine. 2016 Dec;8(1):129.

23. Prüfer K, Muetzel B, Do HH, Weiss G, Khaitovich P, Rahm E, et al. FUNC: A Package for Detecting Significant Associations between Gene Sets and Ontological Annotations. BMC Bioinformatics. 2007 Feb;8:41.

24. Hukku A, Quick C, Luca F, Pique-Regi R, Wen X. BAGSE: A Bayesian Hierarchical Model Approach for Gene Set Enrichment Analysis. Bioinformatics. 2020 Mar;36(6):1689–1695.

25. Zhang S, Cao J, Kong YM, Scheuermann RH. GO-Bayes: Gene Ontology-based Overrepresentation Analysis Using a Bayesian Approach. Bioinformatics. 2010 Apr;26(7):905–911.

26. Weiner 3rd J, Domaszewska T. Tmod: An R Package for General and Multivariate Enrichment Analysis. PeerJ Preprints; 2016.

27. Tiong KL, Yeang CH. MGSEA – a Multivariate Gene Set Enrichment Analysis. BMC Bioinformatics. 2019 Mar;20(1):145.

28. Greene D, Richardson S, Turro E. Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases. The American Journal of Human Genetics. 2016 Mar;98(3):490–499.

29. Hoehndorf R, Hancock JM, Hardy NW, Mallon AM, Schofield PN, Gkoutos GV. Analyzing Gene Expression Data in Mice with the Neuro Behavior Ontology. Mammalian genome. 2014;25(1):32–40.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

30. Davies HTO, Crombie IK, Tavakoli M. When Can Odds Ratios Mislead? BMJ : British Medical Journal. 1998 Mar;316(7136):989–991.

31. Tenny S, Hoffman MR. Relative Risk. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2022. .

32. Harispe S, Ranwez S, Janaqi S, Montmain J. The Semantic Measures Library and Toolkit: Fast Computation of Semantic Similarity and Relatedness Using Biomedical Ontologies. Bioinformatics. 2014 Mar;30(5):740–742.

33. Zhou Z, Wang Y, Gu J. A New Model of Information Content for Semantic Similarity in WordNet. In: 2008 Second International Conference on Future Generation Communication and Networking Symposia. vol. 3. IEEE; 2008. p. 85–89.

34. Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. arXiv preprint cmp-lg/9511007. 1995;.

35. Pesquita C, Faria D, Bastos H, Falcao A, Couto F. Evaluating GO-based Semantic Similarity Measures. In: Proc. 10th Annual Bio-Ontologies Meeting. vol. 37. Citeseer; 2007. p. 38.

36. Horridge M, Bechhofer S. The OWL API: A Java API for OWL Ontologies. Semantic Web. 2011 Jan;2(1):11–21.

37. Haimes Y. On a bicriterion formulation of the problems of integrated system identification and system optimization. IEEE transactions on systems, man, and cybernetics. 1971;1(3):296–297.

38. Phipson B, Smyth GK. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. Statistical Applications in Genetics and Molecular Biology. 2010;9:Article39.

39. North BV, Curtis D, Sham PC. A Note on the Calculation of Empirical P Values from Monte Carlo Procedures. American Journal of Human Genetics. 2002 Aug;71(2):439–441.

40. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. The Human Phenotype Ontology Project: Linking Molecular Biology and Disease through Phenotype Data. Nucleic Acids Research. 2014 Jan;42(Database issue):D966–D974.

41. Johnson AEW, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a Freely Accessible Critical Care Database. Scientific Data. 2016 May;3(1):1–9.

42. Robinson PN. 384 Phenopackets. Zenodo; 2020.

43. Paparoupa M, Spineli L, Framke T, Ho H, Schuppert F, Gillissen A. Pulmonary Embolism in Pneumonia: Still a Diagnostic Challenge? Results of a Case-Control Study in 100 Patients. Disease Markers. 2016;2016:1–8. Available from: `https://doi.org/10.1155/2016/8682506`.

44. Bělohlávek J, Dytrych V, Linhart A. Pulmonary embolism, part I: Epidemiology, risk factors and risk stratification, pathophysiology, clinical presentation, diagnosis and nonthrombotic pulmonary embolism [Journal Article]. Exp Clin Cardiol. 2013;18(2):129–38.

45. Slater LT, Bradlow W, Hoehndorf R, Motti DF, Ball S, Gkoutos GV. Komenti: A Semantic Text Mining Framework. bioRxiv. 2020 Aug;p. 2020.08.04.233049.

46. Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations; 2014. p. 55–60.

47. Bhoj EJ, Li D, Harr M, Edvardson S, Elpeleg O, Chisholm E, et al. Mutations in TBCK, Encoding TBC1-Domain-Containing Kinase, Lead to a Recognizable Syndrome of Intellectual Disability and Hypotonia. American Journal of Human Genetics. 2016 Apr;98(4):782–788.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.14.448423; this version posted August 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

48. Chong JX, Caputo V, Phelps IG, Stella L, Worgan L, Dempsey JC, et al. Recessive Inactivating Mutations in TBCK, Encoding a Rab GTPase-Activating Protein, Cause Severe Infantile Syndromic Encephalopathy. American Journal of Human Genetics. 2016 Apr;98(4):772–781.

49. Guerreiro RJ, Brown R, Dian D, de Goede C, Bras J, Mole SE. Mutation of TBCK Causes a Rare Recessive Developmental Disorder. Neurology Genetics. 2016 Jun;2(3):e76.

50. Zapata-Aldana E, Kim DD, Remtulla S, Prasad C, Nguyen CT, Campbell C. Further Delineation of TBCK - Infantile Hypotonia with Psychomotor Retardation and Characteristic Facies Type 3. European Journal of Medical Genetics. 2019 Apr;62(4):273–277.

51. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The Human Phenotype Ontology in 2021. Nucleic Acids Research. 2021 Jan;49(D1):D1207–D1217.

52. Alazami AM, Patel N, Shamseldin HE, Anazi S, Al-Dosari MS, Alzahrani F, et al. Accelerating Novel Candidate Gene Discovery in Neurogenetic Disorders via Whole-Exome Sequencing of Prescreened Multiplex Consanguineous Families. Cell Reports. 2015 Jan;10(2):148–161.

53. Havrilla J, Zhao M, Liu C, Weng C, Helbig I, Bhoj E, et al. Clinical Phenotypic Spectrum of 4095 Individuals with Down Syndrome from Text Mining of Electronic Health Records. Genes. 2021 Jul;12:1159.