Strategies and techniques for quality control and semantic enrichment with multimodal data: a case study in colorectal cancer with eHDPrep


Published in:
Gigascience

Document Version:
Publisher's PDF, also known as Version of record

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2023 The Authors. This is an open access article published under a Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the author and source are cited.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Open Access
This research has been made openly available by Queen's academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: http://go.qub.ac.uk/oa-feedback
Strategies and techniques for quality control and semantic enrichment with multimodal data: a case study in colorectal cancer with eHDPrep

Tom M. Toner, Rashi Pancholi, Paul Miller, Thorsten Forster, Helen G. Coleman, and Ian M. Overton

Background: Integration of data from multiple domains can greatly enhance the quality and applicability of knowledge generated in analysis workflows. However, working with health data is challenging, requiring careful preparation in order to support meaningful interpretation and robust results. Ontologies encapsulate relationships between variables that can enrich the semantic content of health datasets to enhance interpretability and inform downstream analyses.

Findings: We developed an R package for electronic health data preparation, “eHDPrep,” demonstrated upon a multimodal colorectal cancer dataset (661 patients, 155 variables; Colo-661); a further demonstrator is taken from The Cancer Genome Atlas (459 patients, 94 variables; TCGA-COAD). eHDPrep offers user-friendly methods for quality control, including internal consistency checking and redundancy removal with information-theoretic variable merging. Semantic enrichment functionality is provided, enabling generation of new informative “meta-variables” according to ontological common ancestry between variables, demonstrated with SNOMED CT and the Gene Ontology in the current study. eHDPrep also facilitates numerical encoding, variable extraction from free text, completeness analysis, and user review of modifications to the dataset.

Conclusions: eHDPrep provides effective tools to assess and enhance data quality, laying the foundation for robust performance and interpretability in downstream analyses. Application to multimodal colorectal cancer datasets resulted in improved data quality, structuring, and robust encoding, as well as enhanced semantic information. We make eHDPrep available as an R package from CRAN (https://cran.r-project.org/package=eHDPrep) and GitHub (https://github.com/overton-group/eHDPrep).

Keywords: quality control, semantic enrichment, ontology, colorectal cancer, health data, medical informatics, quality assessment, data integration, bioinformatics
QC may require up to approximately 80% of a data mining project’s time [18]. While data quality and encoding issues in multimodal data can currently be tackled by combining multiple existing approaches, each requires time-consuming familiarisation and may require multiple data transformations, potentially adversely impacting data quality [4]. We present a toolkit for electronic health data preparation (eHDPrep), enabling robust programmatic QC and enrichment of semantic content; high-level functions empower general R users to assess, process, and review their dataset with minimal coding while low-level functions allow advanced R users to specify parameters and workflows as required. We demonstrate the utility of eHDPrep on a multimodal dataset containing 155 variables for 661 colorectal cancer (CRC) patients (Colo-661) [19, 20] and a clinical dataset from The Cancer Genome Atlas (TCGA-COAD) containing 94 variables and 459 colon adenocarcinoma patients. CRC has a large disease burden as the third most common malignancy, with an estimated 1.9 million new cases and 915,800 deaths worldwide in 2020 [21]; advances in CRC medicine are urgently needed [22].

**Findings**

**Quality control**

Data reliability encompasses completeness, consistency, accuracy, uniqueness, and validity [3, 4]. eHDPrep addresses issues in these dimensions through both specific low-level functions and in the high-level functions “assess_quality,” “apply_quality_ctrl,” and “review_quality_ctrl.” The QC workflow in eHDPrep provides user-friendly methods to evaluate and address data quality issues (Fig. 1). We present the application of this workflow to Colo-661 in the sections below in order to enhance data reliability, to enable machine interpretability, and to assess the effects of QC operations upon the dataset. We verify the operations of eHDPrep in TCGA-COAD in the final “Findings” subsection.

**Semantic characterisation**

Following data import, semantic characterisation is required in order to determine variable classifications (Supplementary Table S1) along with information provided by the user, for example, regarding data modalities. The semantic characterisation process includes user review of the automated variable type assignments. Correct semantic information is essential for successful application of downstream steps in eHDPrep.

**Natural language processing**

Analysis workflows typically require data in a standardised format, but significant health data are contained within free-text clinical notes [24]. eHDPrep includes user-friendly extraction of information from free text by wrapping Natural Language Processing functionality from quanteda [12] and tm [25] to create variables describing frequently occurring words or groups of nearby words (eHDPrep function “extract_freetext”). Three free-text variables in Colo-661, containing digitised medical notes, were transformed into 11 new structured variables. Of these, 6 variables were generated from family members’ cancer history (recorded in 21% of patients) following manual correction of observed misspellings, expansion of abbreviations, and standardising cancer name to “[cancer location] cancer” (e.g., “melanoma to “skin cancer”). Four of the new structured variables identified the occurrence of cancer in close family members (mother, brother, sister, father), and 2 further variables recorded if a family member had lung or breast cancer. Manual review determined that the data extraction for these variables had 89.9% sensitivity and 99.9% specificity across the generated values. False positives and negatives were manually corrected in Colo-661.

**Encoding missing values**

Proper representation of missing values is critically important for the correct execution of downstream functions, for example, if missing values are to be excluded from calculations. Missing values may be encoded in a variety of ways, including strings (e.g., “missing,” “unknown”) or out-of-range values (e.g., “−1”) [26]. Indeed, missing values in Colo-661 were recorded in 8 encodings, representing 4.3% of dataset values, which were converted to “NA” values using eHDPrep.

**Completeness**

The degree to which a dataset is populated with data, rather than missing values, is a vital early measurement in quality assessment. eHDPrep measures both variable and patient record completeness at a whole-dataset scale, visualised across Colo-661 in Supplementary Fig. S1. Patterns of completeness may also be explored with eHDPrep through a binary heatmap, the clusters of missing data in Colo-661 showed good correspondence with different data types demonstrating nonrandom missingness (Fig. 2A). Variables with zero entropy [27] (Equation 1) have the same value across all records and, for example, cannot be used to stratify the cohort. Zero entropy variables therefore have limited utility, even if fully complete, and are flagged by eHDPrep. Four Colo-661 variables were removed due to zero entropy. These quality assessment procedures are achieved using the functions “assess_completeness” and “assess_quality.”

\[
H (X) = \sum_{i=1}^{n} p(x_i) \times \log p(x_i) \tag{1}
\]

Where \( p(x_i) \) is the probability of each element \( x \) occurring in the input vector \( X \).

**Internal consistency**

In order to enable evaluation of internal inconsistencies, eHDPrep assesses user-supplied semantic dependencies between variable pairs. In such dependencies, a value in one variable limits the logically valid values in the other. We designed 49 internal consistency checks for Colo-661 across 15 variables (with some variables present in multiple pairs). The checks, as formatted for use in eHDPrep, are shown in Supplementary Table S2. As expected in real-world data, we found 40 instances of internal inconsistency across 5 variable pairs, demonstrating the value of this automated approach.

There was a conflict between the related variables “N stage” and “number of positive lymph nodes” (Fig 3a). One record had a value of “N2” for the “N stage” variable, however, the “number of positive lymph nodes” value was lower than required for assignment of N2 status according to the staging criteria [28]. Similarly, we identified 3 records where the “number of lymph nodes examined” was fewer than the “number of positive lymph nodes” (Fig 3b). Thirty records contained inconsistencies due to a category mislabelling in relation to tumour budding [29] (“high > 10” instead of “high, > 10”), which was identified when comparing a discretised variable with its corresponding nondiscretised variable. Four records stated that patients did not have a personal history of cancer while identifying that the patient had nonmelanoma skin can-
Figure 1: Overview of eHDPrep workflow. The ordering of steps reflects logical dependencies. Dashed arrows and boxes signify optional steps. Following import, the semantic characteristics of the data are established, missing values are dealt with, and a series of operations may be performed. “Natural Language Processing” is only required if free-text variables are present. “Merge Variables” is an optional step for user-defined merging operations with functionality to measure information loss. Variables are encoded in a machine-interpretable format, and a summary report is generated for review by the user. Additionally, functionality to review each step is provided. “Semantic Enrichment” optionally involves aggregation of variables according to semantic commonalities identified by an ontology such as SNOMED CT [23].

Two records showed that the patient had a hereditary form of CRC while stating that the patient had no or an unknown family history of CRC.

Flagging the above inconsistencies focused further data cura-
tion in order to resolve these conflicts. In the above instances, we removed any inconsistent values from one variable in the pair, selected by assessing the reliability of the data source. A more con-
servative strategy might be required if expert curation is not pos-
sible, for example, involving elimination of all conflicting values or potentially removing the inconsistent variables entirely.

**Variable merging**

Merging variables can improve analysis by reducing redundancy and improving storage efficiency. However, inappropriate merging may lead to information loss. Accordingly, we developed functionality in eHDPrep for quantitative evaluation of merging operations using an information-theoretic approach. Information content (IC; Equation 2) is determined from category probabilities for discrete variables or with variable bandwidth kernel density estimation for continuous variables [30]. The mutual information content (MIC; Equation 3) of each input variable with the merged variable is also calculated [30, 31]. Potential information loss during variable merging can be assessed by comparing the MIC of an input variable and the merged variable against the IC of the input variable. If the MIC and IC are identical, the input variable’s information is retained within the merged variable.

\[ IC(X) = - \sum_{i=1}^{n} \log p(x_i) \]  

(2)

\[ MIC = MI(X; Y) \times n \]  

(3)

Where \( p(x_i) \) is the probability of each element \( x \) occurring in the input vector \( X \).

As an example of support for variable merging in eHDPrep, Fig. 4 visualises 2 candidate merging operations applied to Colo-661 variables describing a scoring of Crohn’s-like lymphoid reaction in the tumour [32] based on Graham–Appelman criteria [33]. One possible merging strategy aggregates values “1” and “2” from input1 to “1–2” in the merged variable, leading to information loss (Fig. 4A). A superior strategy takes the values “1–2” from input2 as an ordinal category value between “1” and “2” and does not produce any information loss (Fig. 4B).

**Encode values**

A total of 123 structured variables in Colo-661 were numerically encoded for enhanced machine interpretability. Twenty-two ordinal and 69 binary category variables were encoded as ordered factors, allowing numerical representation of ordinal relationships between values while preserving the original labelling. For example, tumour N stages [28] N0, N1, and N2 were encoded as 1, 2, and 3. Nineteen variables measuring single-substitution mutations were also encoded as ordered factors where the order was determined by the frequency of the mutation status in the cohort; the most common status was encoded as "0" and least common status encoded as "2." Thirteen nonbinary nominal variables were transformed into binary variables describing the presence of
Figure 2: Characterising and comparative measures of completeness in Colo-661. (A) Binary heatmaps showing Colo-661 variable (x-axis) and patient record (y-axis) completeness before QC (left) and after QC (right). Nonnumeric variables were numerically encoded by the number of unique values they possessed. Missing values were numerically encoded as a highly distant value. The dendrograms reflect hierarchical clustering of the values’ Euclidean distances using single-linkage clustering agglomeration. Yellow cells represent missing values, whereas blue cells represent present values. The user-defined data modalities are identified in the coloured bar along the top of each heatmap with a corresponding legend on the right. Multiple regularly shaped areas of missingness are visible, and each block of missingness generally contains variables of the same type. (B) A comparison of variable completeness in the dataset as received (blue) and the cleaned dataset (orange). QC improved overall completeness with a substantial increase in variables possessing >95% completeness. No variables in the cleaned dataset have <28% completeness.

Quality review
Understanding the effect of QC operations applied across large health datasets is nontrivial; eHDPrep simplifies this process and concisely records data changes resulting from QC. First, eHDPrep can produce a comparative tally of unique combinations of values in variables before and after a change was im-
Figure 3: Automated identification of internal inconsistencies with eHDPrep. (A) A boxplot comparing values for “number of positive lymph nodes” (y-axis) and “N stage” (x-axis) in Colo-661. Boxes span the interquartile range with an internal line for the median; whiskers represent the maximum value of the data within 1.5 times the interquartile range beyond the 75th or 25th percentile for the upper or lower whiskers, respectively. Dashed lines show bounds on the “number of positive lymph nodes” for classification into the “N stage” category N1. Some records are inconsistent because they have “N stage” value N2 while the “number of positive lymph nodes” indicates “N stage” of N1. (B) The percentage of positive lymph nodes is shown, derived from analysis of records for the “number of positive lymph nodes” against the “number of lymph nodes examined.” Where values exceed 100% (n = 3), there is a logical inconsistency, because the “number of positive lymph nodes” should not exceed the “number of lymph nodes examined.”

Figure 4: Information-theoretic evaluation of merging operations. A comparison of 2 candidate merging approaches for variables pertaining to Crohn’s-like lymphoid reaction in the tumour. (A) Merging through aggregation of the input 1 values “1” and “2” to “1–2” to align with input 2’s “1–2” value. The mutual information content (MIC) of input 2 with the merged variable is equal to the information content (IC) of input 2; hence, all of the information from input 2 is captured by the merged variable. In contrast, input 1 has IC higher than its MIC with the merged variable and so information loss has occurred. (B) Lossless merging operation where “1–2” values of input 2 were encoded as an intermediate ordinal category between “1” and “2” from input 1. All information from both input variables is contained in the merged variable (i.e., IC is equal to MIC with the merged variable). Indeed, the IC of the merged variable in (B) is greater than the value shown in (A). Therefore, the merging operation shown in (B) is advantageous.

The implemented. These tallies can be produced after each QC action, showing the incremental changes to the dataset. Second, eHDPrep facilitates final review of changes to variable count, which measured 207 variables in Colo-661 following QC, or value-level QC modifications that are optionally summarised in a bar plot. This plot (Fig. 5) highlights differences in the proportion of values modified across Colo-661 that may inform upon the underlying structure of the dataset. Third, eHDPrep’s comparative completeness function visualises the distribution of variable or row completeness before and after QC. Figure 2B demonstrates the positive impact of eHDPrep QC on Colo-661’s variable completeness, resulting in 62% more variables with >95% completeness. Overall, mean variable completeness was 9.45% higher in Colo-661 following QC when compared with the original dataset (following missing value encoding, described above).
Figure 5: Uneven distribution of QC value modifications in the Colo-661 cohort. Stacked bar plot presenting modifications during QC as a percentage of total values per patient is shown on the y-axis. Patient records are displayed on the x-axis, ordered by y-axis values. The proportion of substitutions shows remarkable consistency across most patient records, due to the standardisation of mutation variables that altered all present values, although some patients had missing mutation values that could not be standardised and therefore do not appear.

Semantic enrichment

Ontologies contain valuable curated information about the relationships between domain concepts. Indeed, ontological relationships have been widely employed to support the interpretation of results from high-throughput technologies [34]. Rectangular health datasets (i.e., data frames or matrices) are semantically disorganised. However, ontologies can be utilised to capture semantic relationships between variables during data preparation in a process termed here as “semantic enrichment.” The structure of the ontology provides for aggregation values from constituent variables, in order to generate “meta-variables.” The process is explained below, as applied using the “semantic_enrichment” function in eHDPrep: it is summarised in Fig. 6, and a worked example is provided (Supplementary Fig. S8).

Discovery of the most informative common ancestor terms and semantic aggregation

The IC of each node in the supplied ontology is initially computed to quantify the specificity of nodes by depth and relative number of descendants [35]. Nodes representing variables, manually mapped to the ontology, are added to form an ontology-variable network. Sets of variables that share semantic commonalities through common ontological ancestors are identified. Sets of variables may have multiple common ancestors; therefore, the IC of all common ancestors of a set is compared to identify the most informative common ancestor (MICA), which labels the set. The IC of a MICA can be used to evaluate the semantic similarity of its variable set [36]. Min–max normalisation (Equation 4) is applied to each variable prior to semantic aggregation whereby meta-variables are produced by taking the row-wise sum, minimum, maximum, average, and product of the set for each MICA. Only meta-variables with nonzero entropy (Equation 1) are appended to the dataset.

\[
x_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{4}
\]

Where \( X \) is a numeric vector.

Preparation of ontology graphs

Semantic enrichment in eHDPrep requires 2 edge tables: one describing the mapping of dataset variables to ontological entities and one describing the ontology’s classes and superclasses (child, parent relationships). Generation of the ontology edge table may require preparatory steps, outlined below for our case study with the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [23] and the Gene Ontology (GO) [37, 38].

Variables in Colo-661 were mapped to 2 ontologies (Supplementary Table S1): SNOMED CT, which standardises clinical terms for generation of electronic health records, covering >350,000 concepts at present [23], and GO, a widely used knowledgebase for gene function that currently contains >44,000 terms [37, 38]. The Colo-661 variables were mapped to SNOMED CT by manual review, assisted by the UK National Health Service Digital SNOMED CT Browser [39], producing the first edge table in CSV format. The UK edition of the SNOMED CT Clinical Edition ontology, version 31.1.0, was downloaded from the UK National Health Service’s technology reference data distribution resource [40]. SNOMED CT was converted from Release Format 2 (RF2) to W3C Web Ontology Language (OWL) format using version 2.9.0 of the official SNOMED CT OWL toolkit [41]. We used ROBOT [42] to process and convert SNOMED from OWL to comma-separated values containing each node’s superclasses (Supplementary Fig. S5), enabling generation of the second edge table. Functionality in eHDPrep generated a joined graph from the 2 edge tables. To create the first edge table for GO, variables with gene assignments within the Colo-661 resource were verified and mapped to GO terms accessing Ensembl release 105 [43, 44] with the BiomaRt package [45, 46]. The ontologyIndex package [15] was used to import the January 2022 GO release as the second edge table, which was subset to include only terms from the Biological Process domain and “is_a” ontological ancestors of mapped genes. eHDPrep functionality joined the 2 edge tables in preparation for semantic enrichment.

Enrichment outcomes

A total of 193 (93.2%) of the post-QC Colo-661 variables were mapped to SNOMED CT or GO (Table 2, Supplementary Table S1).
The remaining variables represented negative findings or missing values generated during one-hot encoding, which do not have equivalent entities in SNOMED CT (Supplementary Table S1) and would negate corresponding positive findings if mapped to their entities. Such variables were therefore not mapped and omitted from semantic enrichment, for example, a finding of no diabetes mellitus ("dm_type_NoDM") or missing data for marital status ("maritalcat_NA"). In total, 1,600 meta-variables were generated and appended from 394 variable sets. Approximately 7 times more variable sets were identified per mapped variable in GO than in SNOMED CT with Colo-661, likely due to the 23.2 times greater mean number of distinct annotations in GO that resulted in more common ancestors between variables (Table 1, Supplementary Table S1, Supplementary Table S3). The mean completeness of the Colo-661 meta-variables (98.7%) was 5.1% higher than their constituent variables.

The benefit of semantic enrichment is further demonstrated by 88.2% nonredundant information between the meta-variables from semantic aggregation and their constituent variables in Colo-661, measured by symmetric uncertainty (Equation 5) [47]. Despite low mean redundancy, 9 meta-variables (0.6%) were fully redundant with a constituent variable. For example, for 1 MICA ("beta adrenergic receptor blocking agent therapy", SNOMED CT ID 439630003), 2 of its resultant meta-variables were fully redundant with one of the MICA’s 2 constituent variables, "bisoprolol_cat" (Supplementary Table S4). The redundancy arose from the low number of constituent variables and the low number of unique values that could be taken (0 or 1), which inhibited creation of new information.

$$SU(X, Y) = 2 \frac{H(X) + H(Y) - H(X, Y)}{H(X) + H(Y)}$$  

(5)

Where $X$ and $Y$ are numeric vectors and where $H$ is entropy.

The SNOMED CT network, filtered to only include entities mapped to Colo-661 variables and their ancestors, was found to span many medical subdomains. A selection of these domains is highlighted in Supplementary Fig. S3, ranging from medical procedures, to substances, and to diseases. Selected SNOMED CT MICAs and their descendants are shown in Fig. 7A and Supplementary Figs. S4 and S5. The GO network was similarly filtered and connected several distinct biological concepts (Supplementary Fig. S6) with example GO MICAs shown in Fig. 7B and Supplementary Fig. S7. In Fig. 7A, 5 comorbidity variables are linked by their semantic commonality as types of heart disease while 3 mutation variables are linked by their involvement in drug catabolic processes in Fig. 7B. Figure 7A also visualises another MICA, “ischaemic heart disease,” for 2 variables in the figure. Supplementary Fig. S4 visualises the linkage of 3 variables, describing prescription medications and an adjuvant treatment regimen, as enzyme inhibitor products. The semantic commonalities between 13 variables describing tumour excision location, other surgery types, and emergency surgery status are shown in Supplementary Fig. S5 with the “surgical procedure” MICA. This figure contains additional MICAs for subsets of shown variables, for example, node D ("right colectomy"). Supplementary Fig. S7 shows commonality in a typical feature of several cancers, “negative regulation of programmed cell death” [48], between 16 variables from multiple modalities.

**Comparison with other ontology-focused tools**

Several R packages are available for investigation of variable similarity and to identify common ancestry of ontology terms. We present a qualitative comparison of these tools (Table 2), finding that eHDPrep is unique in generating meta-variables from semantic commonalities.

**Quality control and semantic enrichment of colon adenocarcinoma data from the cancer genome atlas**

eHDPrep was applied to clinical supplement data of colon adenocarcinoma patients from TCGA-COAD [49], further demonstrating the QC and semantic enrichment functionality of eHDPrep. Eight string values representing missing values were standardised to "NA", altering 22,607 values (54.12%). Three instances of internal inconsistency were identified between the recorded American Joint Committee on Cancer (AJCC) pathologic stage and AJCC tumour pathologic T stage. One patient (barcode: TCGA-CK-6746) was assigned a pathologic stage of IIB but a pathologic T stage of T4B. Two other patients (barcodes: TCGA-D5-5539, TCGA-D5-6922) were assigned a tumour stage of IIIA but a pathologic T stage of T3. Thirty-five variables were found to have zero entropy and were removed from the dataset. The ordinal relationships between values in 10 variables were encoded as ordinal factors. Four nominal variables were converted to numeric format with one-hot encoding. Twenty-five binary nominal variables were encoded as factors. Mean variable completeness in the dataset prior to QC was 100% since all missingness was masked by the 8 strings representing missing values. True completeness was found to be 45.88% when strings representing missing values were standardised with eHDPrep. Mean variable completeness was 84.2% following QC, visualised in Fig. 8. Overall, 41 variables were removed and 52 variables...
were added throughout QC. All variables were numeric at the end of QC in contrast to 18.6% before processing with eHDPrep. 

SNOMED-CT was also applied to the TCGA-COAD dataset for semantic enrichment; 98.01% (99 of 101) of post-QC variables were mapped to entities in SNOMED-CT. Thirty-six common ancestors were identified between validation dataset variables, producing 149 meta-variables that were added to the dataset; a further 31 meta-variables were not added because they had zero entropy. The benefit of semantic enrichment is further validated with 85.3% nonredundant information between the meta-variables from semantic aggregation and their constituent variables in TCGA-COAD, measured by symmetric uncertainty (Equation 5). Data and scripts associated with the application of eHDPrep to TCGA-COAD are available in the Supplementary Data File.

**Methods**

**Datasets**

We utilised multimodal data for a cohort of 661 patients as an exemplar for our approach, representing 89% of stage II/III colon adenocarcinoma patients who underwent surgery in 2 healthcare trusts in Northern Ireland between 2004 and 2008 (Colo-661). The cohort primarily comprised colon cancer patients and 2 rectal cancer patients who did not undergo neoadjuvant therapy. Follow-up concluded at the end of 2013, when 212 patients had died from a CRC-specific cause. Subsets of Colo-661 have been used in previous publications and the data had therefore undergone prior manual QC operations [19, 20, 50–56]. Researcher-defined data modalities span clinical pathology, epidemiology, mutation, and treatment and outcomes. These modalities originated from several sources: electronic health records, pathology reports, medical charts, the Northern Ireland Clinical Oncology Information System, Northern Ireland Registrar General’s Office, tumour image analysis [51], ColoCarta panel [57], and targeted mutation analysis. An additional demonstrator dataset was taken from The Cancer Genome Atlas [49], consisting of 459 colon adenocarcinoma patients.

**Data processing**

A combination of base R [58] and the Tidyverse family of packages [13] is used for data formatting and manipulation in eHDPrep. The quanteda package [12] is applied for parsing free-text variables into computable units (tokens), token cleaning, and ultimately to extract information by identifying words that often appear near each other using quanteda’s “tokens_skipgrams” function. The “stopwords” function from tm [25, 59] identifies stopwords, such as “a,” “an,” and “if,” for removal during cleaning. Cluster analysis is implemented through the dist and hclust functions in the stats package [58] to calculate Euclidean distance and single-linkage clusters, respectively, to analyse dataset completeness. Networks for semantic enrichment are created, manipulated, and analysed using igraph and Tidygraph [60, 61].

**Data reporting and visualisation**

The specificity and sensitivity of variables generated from free text in Colo-661, following preprocessing, were assessed through comparative manual review between each value in the generated variables (n = 3,305) and corresponding value in the source variable. Two of the 3,305 values were false positives (99.9% specificity) and 23 were false negatives (89.9% sensitivity). eHDPrep produces summary tables generated using the tibble package [62], which are optionally formatted using knitr [63, 64] and kableExtra [65]. Heatmaps of dataset completeness are visualised with pheatmap [66], and the remaining plots are created with ggplot2 [67]. Network visualisations in this article were generated using Rcy3, Cyotoscape, and InkScape [68–70]. Nodes in Supplementary Figs. S3 and S6 were sized by PageRank centrality [71].

**Discussion**

eHDPrep delivers an accessible set of functions, demonstrated here with suggested workflows applied to 2 real-world medical datasets (Colo-661, TCGA-COAD), demonstrably benefiting data quality. Improvements to Colo-661 using eHDPrep included standardisation of 8 strings representing missingness to “NA”, resolution of 40 internal inconsistencies, conversion of free text to 11 new variables, numeric encoding of 123 nominal and ordinal variables, and a 9.45% increase in mean variable completeness. Additionally, we have showcased tools for assessment of data quality in both the input data and the data following QC operations.

eHDPrep provides novel functionality for data preparation in R where meta-variables are created by aggregation using ontological semantic commonalities. The benefit of this semantic enrichment is exemplified in Colo-661 through the creation of 1,600 meta-variables. Furthermore, the mean redundancy of the meta-variables with their constituent variables was 11.8%, demonstrating creation of substantial information that was absent from the input dataset. The added nonredundant information in meta-variables may potentially enable discovery of patterns where the disaggregated data would be too heterogeneous or too sparse to identify meaningful results. We also observed a 5.1% higher mean completeness of the meta-variables relative to their constituent variables. For records where some values in constituent variables are missing, the meta-variables will contain semantic aggregations of the nonmissing values, affording more comprehensive representation in downstream analyses while simultaneously preserving missing values that may be indicative of patient health and background [72, 73]. A further benefit of eHDPrep is interoperability, an important consideration in digital health care [74, 75]. The standardised encoding of the data im-

---

**Table 1:** Summary statistics of ontologies applied to Colo-661 in semantic enrichment. The GO network has a higher mean number of direct annotations per mapped variable than SNOMED CT, which may explain the larger number of meta-variables generated using GO from fewer mapped variables. Of the 207 variables in the encoded Colo-661 dataset following completion of QC, 14 were not mapped to either SNOMED CT or GO. The number of nodes and edges was measured before the addition of mapped variable nodes.

<table>
<thead>
<tr>
<th>Source ontology</th>
<th>Number of nodes</th>
<th>Number of edges</th>
<th>Number of mapped variables (% of total)</th>
<th>Mean number of direct annotations per mapped variable</th>
<th>Number of variable sets</th>
<th>Number of meta-variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNOMED CT</td>
<td>2,448</td>
<td>4,445</td>
<td>160 (77%)</td>
<td>1.3</td>
<td>160</td>
<td>550</td>
</tr>
<tr>
<td>Gene Ontology</td>
<td>1,675</td>
<td>2,938</td>
<td>33 (16%)</td>
<td>30.2</td>
<td>234</td>
<td>1,050</td>
</tr>
</tbody>
</table>

---

See the full text for further details and references.
Figure 7: Exemplar most informative common ancestors (MICAs) and the semantic relationships with their constituent Colo-661 variables. (A) Relationships between comorbidities are identified by semantic enrichment with SNOMED CT. “Heart disease” is the MICA for all constituent variables (blue) shown in this figure, but “ischaemic heart disease” is also a MICA for the Colo-661 variable set containing only the “mi” and “ihd” variables. (B) Three single-nucleotide polymorphism variables are semantically linked in the GO by the MICA “drug catabolic process.” Mutation variables in Colo-661 were selected based on prior association with colon cancer and exogenous exposures of interest. MICAs can help increase the interpretability of their constituent variables by integrating informative context encapsulated in the ontology; the resulting meta-variable may be useful in downstream analyses, for example, in feature selection for precision medicine applications.
Table 2: Comparison of functionality between eHDPrep and other ontology-focused R packages. eHDPrep provides unique functionality to generate meta-variables using semantic commonalities.

<table>
<thead>
<tr>
<th>Software</th>
<th>Target data</th>
<th>Ontology agnostic</th>
<th>Calculates information content</th>
<th>Calculates semantic similarity</th>
<th>Generates meta-variables from semantic commonalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOSE [17]</td>
<td>Disease Ontology terms; gene products</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>GOSemSim [16]</td>
<td>Gene Ontology terms; gene lists</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>ontologyX [15]</td>
<td>Ontologies</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>eHDPrep</td>
<td>Integrated patient data</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Figure 8: Characterising completeness in TCGA-COAD following eHDPrep QC. A binary heatmap showing TCGA-COAD variable (x-axis) and patient record (y-axis) completeness after QC. Missing values were numerically encoded as a highly distant value. The dendrograms reflect hierarchical clustering of the values’ Euclidean distances using single-linkage clustering agglomeration. Yellow cells represent missing values, whereas blue cells represent present values. Completeness before QC (not shown) was 100% due to strings representing missingness masking true completeness; taking account of the missingness represented in these strings, pre-QC completeness was 45.88%.

Table: Comparison of functionality between eHDPrep and other ontology-focused R packages. eHDPrep provides unique functionality to generate meta-variables using semantic commonalities.

<table>
<thead>
<tr>
<th>Software</th>
<th>Target data</th>
<th>Ontology agnostic</th>
<th>Calculates information content</th>
<th>Calculates semantic similarity</th>
<th>Generates meta-variables from semantic commonalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOSE [17]</td>
<td>Disease Ontology terms; gene products</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>GOSemSim [16]</td>
<td>Gene Ontology terms; gene lists</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>ontologyX [15]</td>
<td>Ontologies</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>eHDPrep</td>
<td>Integrated patient data</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

proves syntactic interoperability and streamlines incorporation into larger databases while the meta-variables support semantic interoperability; for example, during data linkage in identifying similar variables across resources with differing degrees of data aggregation.

Semantic enrichment may be widely useful in health data analysis due to the availability of multiple rich, comprehensive ontologies, for example, the Disease Ontology and the Human Phenotype Ontology [76, 77]. The results of semantic enrichment in eHDPrep are critically dependent upon the ontology taken as input and will likely suffer from a degree of annotation bias [78]. Also, mapping variables to ontology terms can be a time-consuming and complex task requiring background knowledge of the variables if their labels are not self-explanatory; these issues may be mitigated by fuzzy string matching [79] and software interfaces focusing on effective concept search [80], for example, the UK National Health Service Digital SNOMED CT Browser [40]. Importantly, variables generated through semantic enrichment might not properly represent the quantitative relationship between their constituent variables due to unusual associations, such as J-curves [81]. Careful variable aggregation may be applied to avoid variation in 2 variables cancelling each other out. Variables describing missingness or negative findings were omitted from mapping and subsequent aggregation in the case study datasets to avoid problematic aggregations. Aggregations are more numerically meaningful and human interpretable when semantically linked variables measure the presence or amount of the concepts to be aggregated rather than some proportion of variables reporting absent or negative findings. The identified semantic relationships may also aid in interpretation of why the variables have a particular association. For example, if opposite J-curves were found for values of the variables in Fig. 7B, their common involvement in drug catabolism might help to understand the pattern of association. Additionally, the semantic commonalities between variables, via meta-variables, can be used to suggest new semantic dependencies to be tested for internal consistency, further ensuring data quality.

The improved data quality, interoperability, and meta-variables generated through semantic enrichment in eHDPrep are expected to provide for greater robustness and added value in downstream analyses of biomedical data, including Colo-661 and TCGA-COAD.
User Documentation and Technical Details

eHDPrep contains short-form documentation for each function, called with `?function name` in R. Long-form documentation, known as a vignette, is also provided to demonstrate QC and semantic enrichment functionality with synthetic example data, R code, and explanatory text. The vignette is created when the package is built and reflects the functionality of the current version. Error and warning handling messages have been included to ensure expected inputs are received and to notify if unexpected outcomes are returned. eHDPrep is written in the R programming language with a codebase size of 4,533 lines of code and 57 unit tests.

Availability of Supporting Source Code and Requirements

Project name: eHDPrep
Project homepage: https://github.com/overton-group/eHDPrep; https://cran.r-project.org/package=eHDPrep
Operating systems: Windows, Mac, and Linux
Programming language: R
Other requirements: R (≥ 3.6.3)
Licence: GPLv3
RRID: SCR_023401
bio.tools ID: ehdprep (https://bio.tools/ehdprep)

Data Availability

Access to the Colo-661 dataset may be requested by contacting the Northern Ireland Biobank [82] (nibiobank@qub.ac.uk; director Prof. J. James). Synthetic demonstrator data are available within the eHDPrep package. The TCGA-COAD data may be obtained from the NIH GDC data portal [83] or by using the scripts in the Supplementary Data File. An archival copy of the code and supporting data is available via the GigaScience database GigaDB [84].

Additional Files

Supplementary Fig. S1. Completeness in Colo-661. This bar plot summarises patient record (red) and variable (lliac) completeness in the unprocessed Colo-661 dataset. Patient records (red) were between 61% and 94% complete while variable completeness (blue) ranged from 0% to 100%.

Supplementary Fig. S2. Preparation of SNOMED CT for semantic enrichment. ROBOT commands are joined with horizontal arrows. The arguments are shown above their corresponding commands, joined with a vertical arrow with applied parameters displayed in brackets. The “extract” command was used to subset the input ontology. The “BOT” method subset all terms in the ontology to entities mapped to Colo-661 variables plus all super-classes and interrelations between super-classes. The “reason” command was used to logically validate and automatically classify the ontology using the reasoner “hermit” with all tautologies removed. The “relax” command was used to relax Equivalence axioms to weaker SubClassOf axioms, which is suitable for semantic enrichment. The “reduce” command removed redundant SubClassOf axioms using the “hermit” reasoner. Finally, the “export” command exported the ontology as comma-separated values for import into R.

Supplementary Fig. S3. SNOMED CT annotation network. Network nodes represent Colo-661 variables (blue), SNOMED CT terms mapped to Colo-661 variables (green), and their ancestor ontology terms (orange). Node size is proportional to PageRank centrality [71]. The large nodes are highly central, representing domains within the network; some of these are labelled with their SNOMED CT term names (red arrows).

Supplementary Fig. S4. Example of a most informative common ancestor (MICA) term “enzyme inhibitor product.” Variables describing 2 medications (“aceicat” and “lansoprazolecat”) and an adjuvant regimen were semantically linked through the SNOMED CT term “enzyme inhibitor product.” Therefore, semantic enrichment of Colo-661 identified a degree of functional similarity across distinct treatment regimes.

Supplementary Fig. S5. SNOMED CT network for most informative common ancestor (MICA) term “surgical procedure.” Colo-661 variables describing the procedure used to excise the primary tumour, or describing other operations, and the emergency status of the patient’s operation are semantically linked via the “surgical procedure” MICA. This MICA encompasses a relatively large number of variables (n = 13) aggregating information across a range of surgical procedures that could be useful in later analyses. This network also includes several other MICAs, corresponding to smaller groupings of Colo-661 variables, such as node D (right colectomy), which aggregates the variables “right hemicolecotomy” and “extended right hemicolecotomy.”

Supplementary Fig. S6. The Colo-661 Gene Ontology (GO) Biological Process annotation network. Network nodes show Colo-661 variables (blue), their mapped genes (green) with associated GO terms (green), and ancestor terms (orange). Node size is proportional to PageRank centrality [71]. Larger nodes have high PageRank centrality and represent domains within the network, and some of these are labelled (red arrows) with their names from GO.

Supplementary Fig. S7. Gene Ontology Biological Process Network for the most informative common ancestor (MICA) term “negative regulation of programmed cell death.” Network nodes show Colo-661 variables (blue), their mapped genes (green) with associated GO terms (green), and their ancestor terms (orange). The “negative regulation of programmed cell death” MICA describes an important step in the progression of many cancers [48], where cells can evade signals that lead to cell death. Additionally, the figure exemplifies aggregation of variables from different data modalities. For example, “hmgcr_tertile” and “rs12919” at the bottom left of the figure are immunohistochemical and single-nucleotide polymorphism variables, respectively.

Supplementary Fig. S8. Worked example of aggregation in semantic enrichment. Five variables from Colo-661 in the network (blue) were mapped to entities in SNOMED CT that share “heart disease” as their most informative common ancestor (MICA). These 5 variables therefore constitute a “set” and are selected from example synthetic data containing these variables’ names (table, bottom left) and are aggregated row-wise. The aggregated variables (table, bottom right) are appended to the synthetic dataset, labelled with the MICA’s name and the corresponding aggregation function.

Supplementary Table S1. List of Colo-661 variables and ontological mappings (provided as a separate file). Variables that were added, preserved, or removed can be identified by the “Presence Post-QC” column. The user-defined variable modality is recorded in the “Modality” column. Data classes, as encoded in R, are given in the “Data Types” columns. Ranges of values are shown in “Value Range in Post-QC Dataset”: no range is given if a variable was removed from the dataset during QC. Our mapping(s) are provided in the columns “Mapped Ontology” and “Ontological/Gene Mapping.” Variables mapped to the GO require an initial mapping to a gene, shown here, with Gene GO term mappings detailed in Supplementary Table S3.
Supplementary Table S2. Internal consistency checks performed on Colo-661 (provided as a separate file). Tests between variables both containing numeric values were performed using the logical operator in "Logical Test" with the format "[Variable A] [Operator] [Variable B]." Between variables containing categorical values, values in "Variable B Boundaries" were tested to only be present given the corresponding values in "Variable A Boundaries." Tests between numeric and categorical values were similarly compared with numeric (inclusive) ranges denoted by colon-separated values.

Supplementary Table S3. List of genes mapped to Colo-661 variables and the mapped GO terms (provided as a separate file). GO terms are separated by ",". Mappings between genes and GO terms were sourced from Ensembl.

Supplementary Table S4. Redundancy between meta-variables and a constituent variable. Two of 5 meta-variables derived from minimum and product aggregations of "bisoprolol_cat" and "betablocker_cat" (semantically linked by the MICA: "beta adrenergic receptor blocking agent therapy") were fully redundant with "bisoprolol_cat." The table describes the observed row-wise combinations of values across the constituent variables and the 2 meta-variables that were redundant with "bisoprolol_cat." While the value of "betablocker_cat" differed from values of the meta-variables shown, the value of "bisoprolol_cat" did not, which led to the observed redundancy. "bisoprolol_cat" and "betablocker_cat" described if patients were prescribed bisoprolol and beta-blockers, respectively.

Abbreviations
AJCC: American Joint Committee on Cancer; COAD: colon adenocarcinoma; CRC: colorectal cancer; GO: Gene Ontology; IC: information content; MIC: mutual information content; MICA: most informative common ancestor; QC: quality control; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms, TCGA: The Cancer Genome Atlas.

Ethics approval and consent to participate
Provision and use of this dataset was approved by the Epi700 Consortium and Northern Ireland Biobank under Secondary Use of Data. REC references: 10/NIR02/53, 11/NI/0013.

Competing Interests
I.M.O. has provided consultancy for Mevox Ltd for work unrelated to this publication. The authors declare that they have no other potentially competing interests.

Funding
LifeArc (I.M.O., H.G.C.), Engineering and Physical Sciences Research Council (2280988; I.M.O., H.G.C., P.M.). Health Data Research UK (HDR-UK) Substantive Site (I.M.O., H.G.C.); HDR-UK is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation, and Wellcome. The funders had no influence upon study design, collection, analysis or interpretation of data, or in writing the manuscript.

Authors’ Contributions

Authors’ information
T.M.T. is a PhD candidate in the Overton group, Patrick G. Johnston Centre for Cancer Research (PGJCCR), Queen’s University Belfast (QUB). R.P. is a former MSc research project student in the Overton group, PGJCCR, QUB. P.M. is a Reader, theme lead for Cybersecurity, and deputy director of The Centre for Secure Information Technologies, QUB. T.F. is a Principal Data Scientist within LifeArc’s Data Sciences Group and Project Leader for electronic health record projects. H.G.C. is Professor of Cancer Epidemiology and the Cancer Epidemiology research group lead in the Centre for Public Health and PGJCCR, QUB. I.M.O. is a Reader and Data Intensive Biomedicine group lead in the PGJCCR, QUB.

Acknowledgements
We are grateful to the Epi700 Steering Committee for curation of and support with the data resource containing Colo-661. The samples used in this research were received from the Northern Ireland Biobank, which has received support from Health and Social Care Research and Development Division of the Public Health Agency in Northern Ireland and Cancer Research UK (via the former Belfast CRUK Centre and the Northern Ireland Experimental Cancer Medicine Centre) and the Friends of the Cancer Centre. The Northern Ireland Molecular Pathology Laboratory, which was responsible for creating resources for the Northern Ireland Biobank, has received funding from Cancer Research UK, the Friends of the Cancer Centre, and the Sean Crummy Foundation. Clinical data collected and analysed were facilitated by The Northern Ireland Cancer Registry, which is funded by the Public Health Agency, Northern Ireland. Thanks to Dr. Alexander Lubbock and Dr. Seanna McTaggart for helpful discussions, also to the Overton group for testing eHDPrep. Thanks to Dr. Hadley Wickam and Davis Vaughan for identifying and ensuring eHDPrep reverse compatibility with the dplyr [85] R package.

References


