Assessment Cover Sheet for Essay

A signed and completed cover sheet must be securely attached to all work submitted for assessment.

Work submitted without a cover sheet will not be marked.

Student Name:	Steven Mitchell
Student Number:	40175882
Word Count:	2,425

Declaration of academic integrity

I declare that I have read the Queen's University regulations on plagiarism, and that the attached submission is my own original work.

It is not similar in content to, nor based on the work of others, whether published or unpublished, except with full and proper acknowledgement.

I consent to my work being submitted to the plagiarism software used by Queen's University Belfast.

Signed: ________________________________

Date: **28 APR 2023**

1. You have been tasked with the analysis of a health information system and you are told that the system is able to work with codes derived from the International Classification of Diseases (ICD). However, there is no information concerning the revision of the codes the system is using. You are requested to test the system using the code for confirmed Respiratory tuberculosis. Using the latest versions of ICD-10 and ICD-11 (hint use WHO resources) please:

a. Give a detailed description of the tools, strategy and reasoning used to retrieve these codes, provide the codes and exact names that are used for this concept on each of the ICD revisions. (4 points)

Retrieval of the codes for confirmed respiratory tuberculosis in ICD-10 and ICD-11 can be achieved using the following tools and strategy:

Tools: ICD-10 browser https://icd.who.int/browse10/2019/en

ICD-11 browser version 01/2023 https://icd.who.int/browse11/l-m/en

Accessed via the World Health Organization website, these are the latest versions of the International Classification of Diseases (ICD).

Strategy: Search the term ‘Respiratory tuberculosis’ in both browsers, as they may differ.

Results:

ICD-11	
1B10	Tuberculosis of the respiratory system
1B10.0	*Respiratory tuberculosis, confirmed*
1B10.1	*Respiratory tuberculosis, not confirmed*
1B10.Z	*Respiratory tuberculosis, without mention of bacteriological or histological confirmation*

ICD-10	
A15-A19	Tuberculosis
A15	Respiratory tuberculosis, bacteriologically and histologically confirmed
A16	Respiratory tuberculosis, not confirmed bacteriologically or histologically

A15.0-9 & A16.0-.9 denote means of confirmation testing

2. In the context of ICD coding system:

a. Please identify what diseases/conditions are encoded in by 1C42&XT8W and 1C1G/8B88.0 (2 point)

1C42&XT8W: Chronic Melioidosis

1C1G/8B88.0: Lyme borreliosis induced Bell palsy

b. Expand by providing additional details about how these codes were built and the similarities and differences on how they were built and their components (4 points)

Both codes begin with 1 which denotes the diseases fall under chapter 1 of ICD-11 (1A00–1H0Z) "*Certain infectious or parasitic diseases*". The chapter character is followed by a letter, in this case C, which is also common to the two codes indicating "*Other bacterial diseases*". A number and a fourth character (which can be a letter or number) follows. 1C**42** represents Melioidosis whereas 1C**1G** denotes Lyme borreliosis. Cluster coding is then used to show an additional code with / used for a stem code and & for an extension code, which denotes a clinical concept. For the first code a forward slash with XT8W indicates the chronic course and for the second, the ampersand represents "*has manifestation*" with the code 8B88 indicating Bell palsy. The benefit of this cluster system is that 8B88.0 can be traced on its own to find the sole indication of Bell palsy that is not as a result of Lyme disease.

c. State which ICD revision do these codes belong to (2 point). Describe and provide details about the reasoning used to assign these codes to a given ICD revision and describe the main characteristics of this revision (i.e.number of chapters, structure of the coding system…) (3 points)

Both codes belong to ICD-11. ICD-10 chapters begin with a roman numeral 1 through to 22, followed by block numbers of a letter and two numbers, which is unlike the provided codes which belong to the more flexible coding structure of the ICD-11 format wherein there is a single first character for every chapter, with the first nine chapters starting with the numbers 1 to 9 and the next nineteen chapters starting with the letters A to X. Legibility features are built into the format including the omission of the letters I and O to prevent confusion with 1 and 0. ICD-11 allows for a more detailed hierarchy of parent-child relationships. Ultimately the coding system in ICD-11 is more flexible and there is no limitation on the number of subcategories per code. Additionally, more chapters were added to stay in line with modern medicine such as "*Conditions related to Sexual Health*", "*Disorders of the Immune system*" and "*Sleep-wake disorders*". Stem codes were added to allow the combination of two conditions into a single code, instead of a pre-coordinated code, the first more generic code can denote a condition and additional stem code characters can indicate more specific details such as location or behavior.[1]

3. Using the online demo version of gnuhealth HER (http://federation.gnuhealth.org:8000/) (user name: admin; password: gnusolidario) Select patient "Ana Isabel Betz" and answer the following questions.

a. How many conditions does this patient have? (2 points)

Z88.0: Personal history of allergy to penicillin (severe). **E10:** Type 1 diabetes mellitus (chronic). **A90:** Dengue fever [classical dengue] (acute). **A40**: Streptococcal sepsis. ***(not encoded)*** severe allergic reactions to β-lactams

5 conditions shown.

b. Are the conditions encoded in any manner? If encoded, what code is being used? Explain your answer. (3 points).

The conditions are encoded with ICD-10 nomenclature. Encoding is consistent with ICD-10 ie. a letter and two digits, and all condition codes when searched via ICD-10 produce the listed associated condition.

4. In the context of UMLS

a. Using your own words explain what the letter C in C0012813 and A in A0050653 mean and what is the relationship between both of them. (4 points)

C0012813: C is for concept, denoting this is a Concept Unique Identifier (CUI). Concepts are a key component of the UMLS solution; different sources (such as MeSH or ICD-11) have different terminologies for diseases that may not be easily linked. UMLS links synonymous terminology via a coding system; it groups names from different terminologies into units of meaning known as concepts. Every name in UMLS is grouped with its synonyms and this grouping is given a unique, shared number, the CUI.

A0050653: A is for atom, denoting an Atom Unique Identifier (AUI). Each unique string or name (like Alzheimer Disease) from a unique source (like MeSH) with a unique code (like D000544) is called an atom. Each atom gets a unique identifier. All atoms considered synonyms of a single disease are given the same CUI, but each atom has its own unique number. This is how names and codes are linked together from different sources.

b. Using UMLS Terminology Services at the NIH explain what is encoded by both codes above. (2 points)

C0012813: Diverticulitis (concept CUI, Inflammation of a DIVERTICULUM or diverticula)

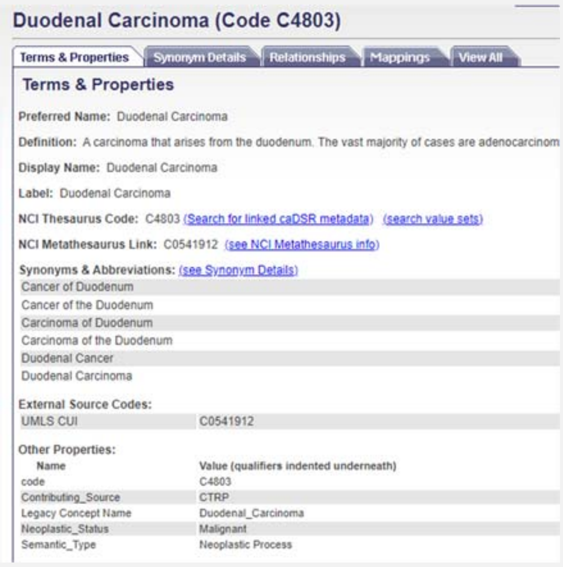
A0050653: Diverticulitis (atom, from the Medical Subject Headings or 'MeSH' code developed and maintained by the U.S. National Library of Medicine, whose MeSH code is D004238)

c. Describe in detail an approach using UMLS that would allow you to identify the equivalent NCI Thesaurus term(s) to the MeSH term D004379. (hint: you can use the UMLS Terminology services to help you) (6 points)

The MeSH term D004379 was searched.via the UMLS Metathesaurus browser. The concepts Duodenal Cancer (C0541912) and Duodenal Neoplasms (C0013291) were returned. Under NCI definitions the thesaurus entry for each could be determined, additional information for NCI atoms could also be determined from the CUI pages for each (showing PT for preferred terms and SY for synonyms). Results could then be verified on the NCI thesaurus browser by searching the code (exact match).

C0541912: Duodenal Cancer. Definition (NCI) *A carcinoma that arises from the duodenum. The vast majority of cases are adenocarcinomas*. Atoms A7593432 (PT '*Duodenal Carcinoma*') A7589513, A7589508, A7589507, A7634189, A7634335. All atoms had NCI code C4803. All atoms had NCI code C2995. NCI entry shown in *figure 1*.

C0013291: Duodenal Neoplasms. Definition (NCI) *A benign or malignant neoplasm that* *affects the wall of the duodenum. Representative examples include adenoma, carcinoma, and lymphoma*. Atoms A7570166 (PT '*Duodenal Neoplasm*') A7593443, A7610285, A7610562, A7627523, A7627805. All atoms had NCI code C2995. NCI entry shown in *figure 2*.



Duodenal Carcinoma (Code C4803)

Terms & Properties

Preferred Name: Duodenal Carcinoma

Definition: A carcinoma that arises from the duodenum. The vast majority of cases are adenocarcinom

Display Name: Duodenal Carcinoma

Label: Duodenal Carcinoma

NCI Thesaurus Code: C4803 (Search for linked caDSR metadata) (search value sets)

NCI Metathesaurus Link: C0541912 (see NCI Metathesaurus info)

Synonyms & Abbreviations: (see Synonym Details)

Cancer of Duodenum
Cancer of the Duodenum
Carcinoma of Duodenum
Carcinoma of the Duodenum
Duodenal Cancer
Duodenal Carcinoma

External Source Codes:

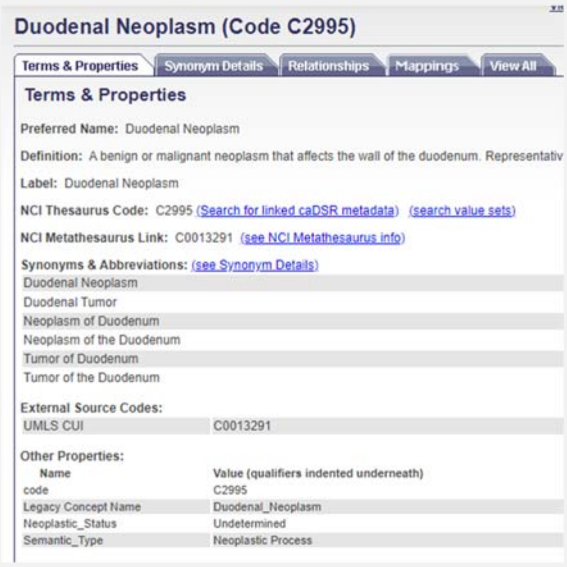
UMLS CUI	C0541912

Other Properties:

Name	Value (qualifiers indented underneath)
code	C4803
Contributing_Source	CTRP
Legacy Concept Name	Duodenal_Carcinoma
Neoplastic_Status	Malignant
Semantic_Type	Neoplastic Process

Fig. 1

NCI entry for C2995, derived from MeSH term D004379



Duodenal Neoplasm (Code C2995)

Terms & Properties

Preferred Name: Duodenal Neoplasm

Definition: A benign or malignant neoplasm that affects the wall of the duodenum. Representativ

Label: Duodenal Neoplasm

NCI Thesaurus Code: C2995 (Search for linked caDSR metadata) (search value sets)

NCI Metathesaurus Link: C0013291 (see NCI Metathesaurus info)

Synonyms & Abbreviations: (see Synonym Details)

Duodenal Neoplasm
Duodenal Tumor
Neoplasm of Duodenum
Neoplasm of the Duodenum
Tumor of Duodenum
Tumor of the Duodenum

External Source Codes:

UMLS CUI	C0013291

Other Properties:

Name	Value (qualifiers indented underneath)
code	C2995
Legacy Concept Name	Duodenal_Neoplasm
Neoplastic_Status	Undetermined
Semantic_Type	Neoplastic Process

Fig. 2

NCI entry for C2995, derived from MeSH term D004379

5. Using NCBO's BioPortal (https://bioportal.bioontology.org/) tools (default parameters) and the abstract for the following article "Coldbeck-Shackley RC, Romeo O, Rosli S, et al. Constitutive expression and distinct properties of IFN-epsilon protectthe female reproductive tract from Zika virus infection. PLoS Pathog. 2023 Mar 10;19(3):e1010843. doi: 10.1371/journal.ppat.1010843. Epub ahead of print. PMID: 36897927.

Please report on the following:

a. How many ontology classes can be identified in this abstract? (2 points). Describe and provide details about the tool and the procedures used to generate the result. (2 points)

Using Annotator, wherein the abstract text was entered into the search box, using default settings, a total of 3132 results where returned. Annotator plus was able to return 1184, which removed many of the repeated classes from the original Annotator tool. Both Annotator tools use NLP algorithms to identify ontology classes in the text, matching text with concepts from various biomedical ontologies and returning a list of annotations with their respective ontology classes.

b. Using the results from above (5a), how many unique ontologies were these classes associated to. Explain how you reach this conclusion. (2 points).

Using an excel countif function it was found 347 unique ontologies were associated with the classes identified by the annotator tool. Data was taken from the tool and imported to excel were it used a basic =COUNTA(UNIQUE(range))-1) formula to find unique ontologies.

c. Describe which tool would you use to decide which ontology would provide a better coverage (identify more terms within this text) and provide the reasoning behind this decision and explain the importance of the different parameters used (how many ontologies, coverage, etc…) (6 points)

To decide which ontology would provide better coverage for the text, the NCBO BioPortal Recommender tool could be used. This tool can recommend ontologies that are most relevant to a given text,and allocates a final score to show which would be best suited. The score takes into account various parameters such as coverage, acceptance, detail and specialization. The coverage score is based on the extent to which the given ontology covers the terms present in the input text, whereas the acceptance score reflects the level of recognition and reliability of the ontology, essentially its providence in the research community. The knowledge detail score evaluates the level of specificity and comprehensiveness of the ontology, such as whether it includes definitions, synonyms, and other often important details for further research. Finally, the specialization score assesses how well the ontology covers the specific domain of the input.[3]

In this case, National Cancer Institute Thesaurus (NCIT) was found to most likely provide the most relevant ontologies for the text with the highest final score of 63.5, followed by SNOMEDCT at 43.8.

6. Medical Subject Heading codes are commonly used in the biomedical domain for multiple purposes including annotation of genes. Using the meshR tool analyse whether the following gene set is enriched in any particular MeSH term. The file with the gene lists is available on Canvas (assignment2223_meshr). Use a significance of adj. pvalue =1e-6 and gene2pubmed annotations. Please explain the following aspects:

a. Are there any MeSH terms (Disease related MeSH terms) that are enriched in this particular dataset? (2 points)

b. Provide the list of the MeSH terms (and codes) that have been found statistically significant, explain the relevance and role of the different hyperparameters used and provide the code required. (8 points).

Hyperparameters are used in meshR to allow analysis to be customized according user requirements, in this case to adjusting the significance level or selecting a different MeSH term type. Annotation specifies the database to be used for gene ID mapping (eg gene2pubmed, and MeSH category specifies the terms to be used for analysis (eg disease). The p value cut off specifies the threshold for the adjusted p-value, below which terms are considered to be significantly enriched. In this case, a threshold of 1e-6 is used.

Code is found in Appendix I

7. Explain briefly in the context of Natural Language Processing:

a. What would be the impact of removing punctuation and numbers during tokenisation in a gene detection pipeline? (3 points)

NLP uses tokenization to break text down text into individual words or tokens. Punctuation is generally treated as individual tokens, as are numbers and total removal could have significant negative effects on a gene detection pipeline wherein hyphens and numbers may frequently be used within gene names, and cause the system to miss genes in text. However, several studies have trained systems on rules that are more bespoke to genes, such as one study[2] that utilized several rules to develop a more tailored gene pipeline friendly system. These rules included the removal of characters rarely found in gene names, the removal of punctuation marks utilized outside gene names (actual punction), the removal of brackets outside gene name, rules to deal with quotations and apostrophes and then finally slashes used outside gene names. Removal of these non-functional characters was able to enhance the efficiency of the pipeline whilst causing minimum impact to accuracy.

b. In the comparison between two systems (A and B) for the detection of genes that have been trained and run on the same dataset, system A had a precision of 0.772 and a recall of 0.931 and system B has a precision of 0.935 and a recall of 0.702.

• Which of these systems returns more genes? Explain your answer. (2 points)

System A has a higher recall value of 0.931 compared to B with 0.702 which could mean because it is able to detect more relevant information in the data, it may be able to return more genes from the dataset. However, it cannot be determined for certain without additional information regarding the total genes in the dataset and number predicted by each dataset.

• Which of these performs better overall? Justify and explain your answer. (4 points)

System A has a lower precision of 0.772 (vs 0.935 for B) but a higher recall score of 0.931 (vs. 0.702).To determine which system performs better the F1 score for each system can be determined wherein F1 score = 2 * (precision * recall) / (precision + recall). For system A the F1 score is 0.844 and for system B it is 0.801. Overall, this indicates system A performs better when balancing precision and recall; it is same to correctly identify more genes whilst maintaining reasonable precision. System B is more precise but also more likely to miss genes in the dataset.

8. Using CTDbase (http://ctdbase.org/) please report on the following aspects providing detailed information about how you retrieved the results:

a. How many genes are associated with Autistic Disorder? (2 points)

Method: A keyword search for "*Autistic Disorder*" was performed filtering for *Diseases*. The *Gene* tab was selected, and results exported to Excel where additional count processing was performed to determine the number of genes associated with Autistic Disorder (COUNTA or COUNTA(UNIQUE) selection formulas).

Result: 25,445 results were exported from the search, 25,441 of which were found to be unique genes associated with Autistic Disorder. 262 were classed as having a marker/mechanism and/or therapeutic association to the disorder.

b. How many chemicals have been associated with Autistic Disorder? How many show a “direct therapeutic evidence”? (2 point)

Method: A keyword search for “*Autistic Disorder*” was performed filtering for *Diseases*. The *Chemical* tab was selected, and results exported to Excel where additional count processing was performed to determine the number of genes associated with Autistic Disorder (COUNTA or COUNTA(UNIQUE) selection formulas).

Result: 7,818 results were exported from the search. 7,077 of these were unique chemicals and 108 were classed as having a marker/mechanism and/or therapeutic association to the disorder, of these 41 where classed only as therapeutic.

c. How many exposure studies have been reported in relation with Autistic Disorder? (1 points)

Method: A keyword search for “*Autistic Disorder*” was performed filtering for *Diseases*. The *Exposure studies* tab was selected and the number of results recorded.

Result: 9 exposure studies were associated with the disease.

d. What chemical(s) was(were) studied as stressor(s) in the study(ies) that used plasma as a sample? (2 points)

Method: A keyword search for “*Autistic Disorder*” was performed filtering for *Diseases*. The *Exposure studies* tab was selected, and results filtered by medium to ‘Plasma’.

Result: Perfluorodecanoic acid, perfluorohexanesulfonic acid, perfluoro-n-nonanoic acid, perfluorooctane sulfonic acid, and perfluorooctanoic acid where studied as stressors.

e. Provide information about how many pathways are significantly enriched (with an adj. p-value of 0.01) for the chemical compound(s) referred above and provide a detailed description on how you generated these results. (5 points)

Method: A keyword search for each chemical compound was performed filtering for *Chemicals*. The Pathways tab was then selected where results with corrected p-value <0.01 where shown, and results recorded.

Result:

Chemical	n enriched (p <= 0.01)
Perfluorodecanoic acid	218
Perfluorohexanesulfonic acid	142
Perfluoro-n-nonanoic acid	401
Perfluorooctane sulfonic acid	3,127
Perfluorooctanoic acid	842

9. Given the following compounds Perfluorooctane sulfonic acid, Oxybenzone and Perfluorobutane and using CTDbase answer the following questions and explain how you achieved these results:

a. Generate a Venn diagram showing how many genes these three compounds have in common and how many are unique? (3 points).

Method: On CTD database, VennViewer under the Analyze tab was selected. Chemicals was selected as the input type and the 3 chemicals Perfluorooctane sulfonic acid, Oxybenzone and Perfluorobutane were entered. Gene associations (curated) was selected. Chemical–gene interaction type was kept at ANY (increases, decreases, affects). Under Hierarchical or Direct Relationships, 'Commonalities between exact terms only' was selected. The form was submitted.

Result: Venn diagram shown in *figure 3*. Perfluorooctane sulfonic acid had 335 unique genes, Oxybenzone had 167 and Perfluorobutane and 1. Oxybenzone and Perfluorobutane had 0 interactions exclusively in common, whereas Perfluorooctane sulfonic acid had 85 with Oxybenzone and 14 with Perfluorobutane. All three chemicals had 4 gene interactions in common, CPT1B, CYP7A1, HADHA and PPARG.

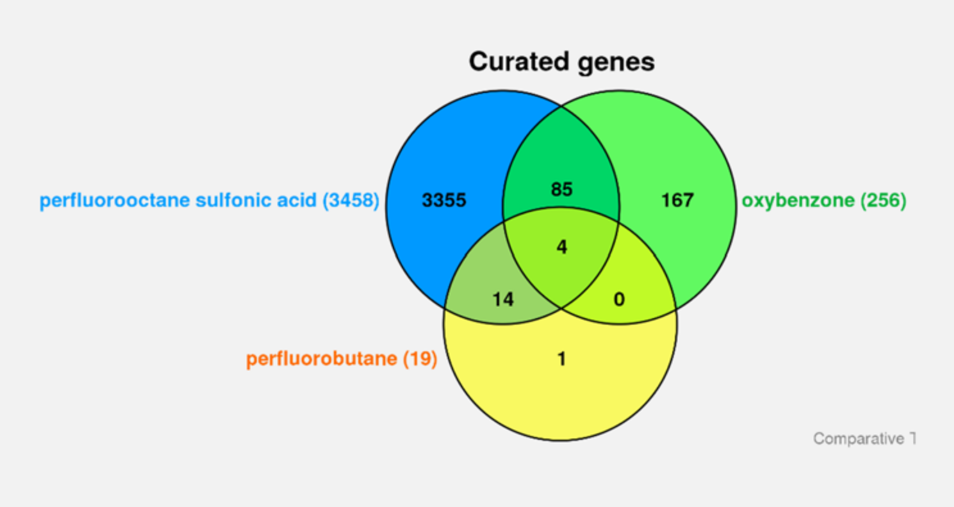


Fig. 3

Curated gene interactions for the chemicals shown in a Venn Diagram

b. Identify the common and unique pathways associated with these chemical compounds using enriched pathways and inferred pathways. Compare the results and provide an explanation about the differences observed. (6 points)

Method: On CTD database, VennViewer under the Analyze tab was selected. Chemicals was selected as the input type and the 3 chemicals Perfluorooctane sulfonic acid, Oxybenzone and Perfluorobutane were entered. Pathway associations > enriched was chosen first, then inferred. The form was submitted. Results where outputted to excel to determine degree of overlap in values (using a COUNTIF function).

Result: 12 enriched pathways are common to all chemicals (*figure 4*), which are predominantly related to fatty acid metabolism and PPAR signaling pathway. Enriched pathways are determined based on actual gene expression changes observed in experimental studies, while inferred pathways are based on predictions of gene expression changes based on known interactions between genes, proteins, and chemicals; they are algorithm based. The interred pathways, of which there are many more (141 as shown in *figure 5*) do include 11 of the 12 enriched pathways. This exposes the possible need for more research, and possible further curation of the inferred database, which extends beyonf metabolism and into cancer pathways and tumour suppression (eg p53 signalling pathway, Prostate cancer, Thyroid cancer).

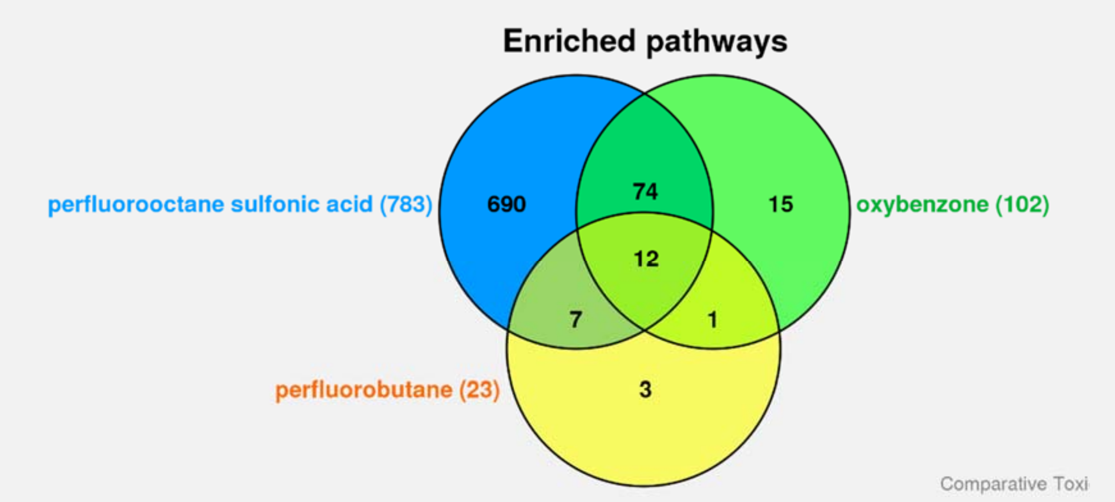


Fig. 4

Enriched pathways associated with Perfluorooctane sulfonic acid, Oxybenzone and Perfluorobutane

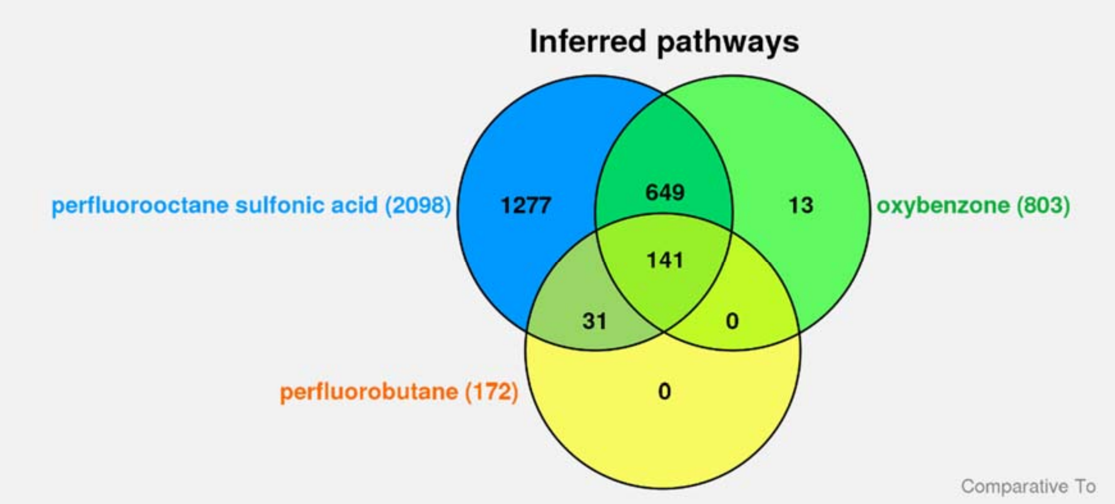


Fig. 5

Inferred pathways associated with Perfluorooctane sulfonic acid, Oxybenzone and Perfluorobutane

10. Using the phexpo package* (https://pubmed.ncbi.nlm.nih.gov/32734156/) answer the following questions.

a. Provide a list of the phenotypes identified for the chemical “3,4,5,3',4'- pentachlorobiphenyl” and explain the results obtained. (2 points)

Based on the Bonferroni-corrected p-value (as shown in *figure 6)*, only the skin nodule phenotype is significantly associated with 3,4,5,3',4'- pentachlorobiphenyl exposure at a significance level of 0.05. However, more phenotypes are significant based on FDR (*figure 7*) including Abnormal coronary artery morphology, Myocardial infarction, Abnormal circulating metabolite concentration, Vomiting, Abnormal cardiovascular system physiology, Nausea and vomiting, and Thick lower lip vermilion. These results suggest that exposure to "3,4,5,3',4'-pentachlorobiphenyl" may be associated with certain phenotypes related to cardiovascular health, as well as gastrointestinal symptoms.Ultimately, further studies would be required to confirm these associations and to investigate the underlying mechanisms.

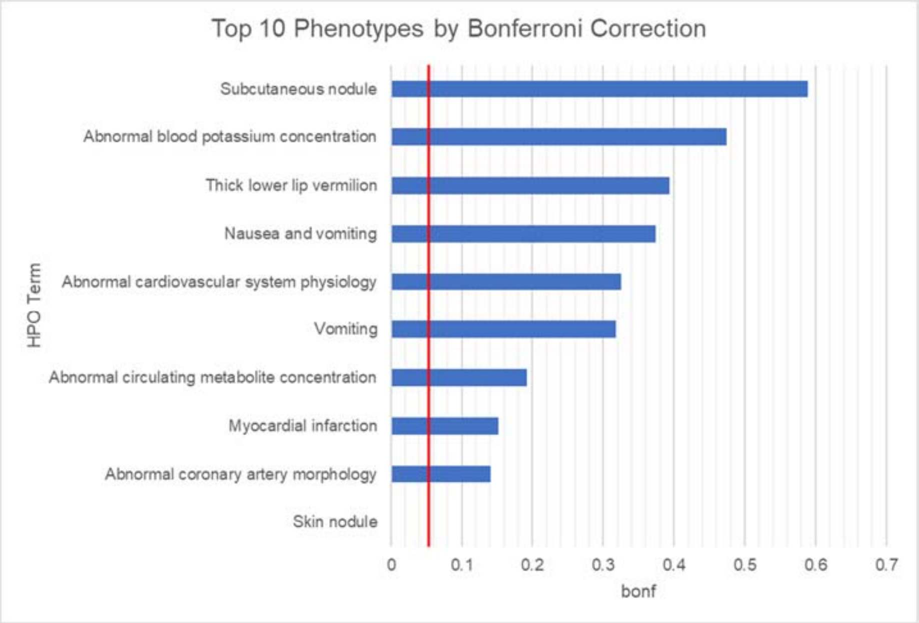


Fig. 6

HPO terms identified for 3,4,5,3',4'- pentachlorobiphenyl filtered by Bonferroni correction with significance threshold at 0.05.

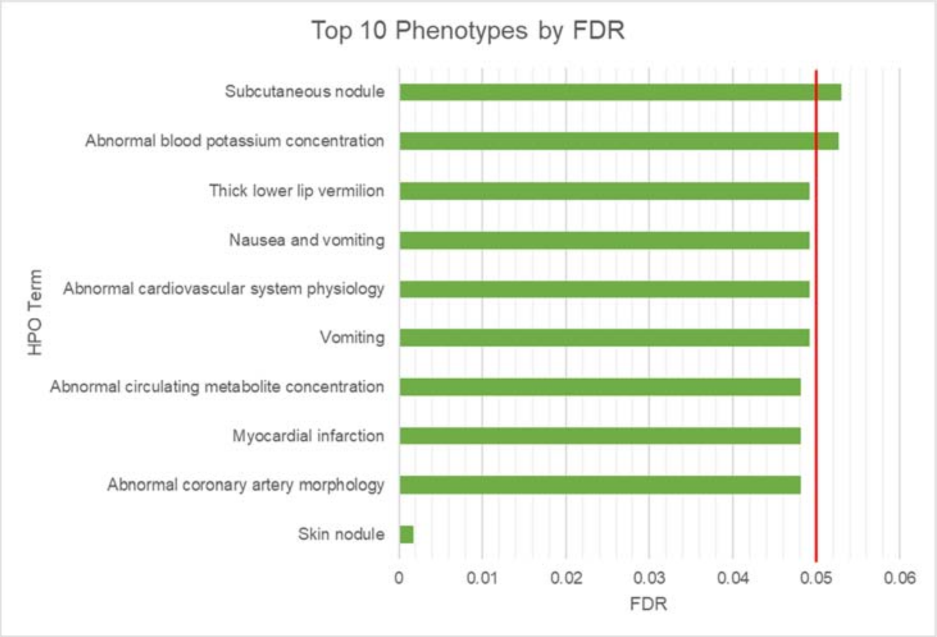


Fig. 7

HPO terms identified for 3,4,5,3',4'- pentachlorobiphenyl filtered by false discovery rate with significance threshold at 0.05.

b. Compare the methodology and results between CTDbase phenotypes associated with this compound and those identified by phexpo and explain similarities and differences. (5 points)

CTDbase and phexpo can be used to identify phenotypes associated with exposure to selected chemicals. CTDbase is a curated database integrating human curated data with algorithm defined inferred results to generate chemical-gene, chemical-disease, and gene-disease interactions.In contras, phexpo is a software plugin that identifies phenotypes associated with exposure to chemicals by analysing electronic health records.In terms of the phenotypes associated with "3,4,5,3',4'-pentachlorobiphenyl", CTDbase identifies 194 phenotypes, including cell population proliferation, membrane lipid catabolic process, increased response to oxidative stress, and effects on cell growth. Disease states linked include Neoplasm Metastasis, Liver Cirrhosis, Atherosclerosis, Fatty Liver and heart disease. Several of these phenotypes are related to cardiovascular health and which is consistent with some of the phenotypes identified by phexpo. However, some conditions identified by CTDbase such as Hypercholesterolemia is not specifically identified by phexpo. Conversely, gastro issues like vomiting and nausea are not called out by CTDbase. The differences in the phenotypes identified by CTDbase and phexpo may be due to the different methodologies used by these tools; where CTDbase relies on curated literature and biological data, phexpo uses electronic health records. Additionally, the different phenotypes identified by these tools may reflect different aspects of the health effects of exposure to "3,4,5,3',4'-pentachlorobiphenyl". Again, further research is needed to reconcile these differences and to gain a better understanding of the health effects of 3,4,5,3',4'-pentachlorobiphenyl.

References

[1] Pickett, D., 2018. Update on ICD-11: The WHO Launch and Implications for U.S. Implementation. [Online] Available at: https://www.cdc.gov/nchs/data/icd/ICD-11-WHOV-CM-2018-V3.pdf [Accessed 24 April 2023].

[2] Jiang, J. & Zhai, C., 2007. An empirical study of tokenization strategies for biomedical information retrieval. Information Retrieval volume , Volume 10, p. 341–363.

[3] Zulkarnain, N. Z., Meziane, F. & Crofts, G., 2016. A Methodology for Biomedical Ontology Reuse. Natural Language Processing and Information Systems, Volume 9612, p. 3–14.

```
## APPENDIX I : CODE


setwd("C:/Users/swmit/OneDrive - Queen's University Belfast/SCM8148 Health and
Biomedical Informatics and the Exposome/Written Report")

# QUESTION 6

# install required packages
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install(version = "3.16")

BiocManager::install(c("meshr", "AnnotationHub", "MeSHDbi"))

      # Load required libraries
      library(meshr)
      library(AnnotationHub)
      library(MeSHDbi)
      library(BiocManager)

  # Read the gene set file
  genes <- read.delim("assignment_meshr.csv", header = TRUE)

  # Database preparation
    ah <- AnnotationHub()
    d <- display(ah)

    dbfile1 <- query(ah, c("MeSHDb", "MeSH.db", "v002"))[[1]]
    dbfileQ <- query(ah, c("MeSHDb", "Homo sapiens", "v002"))[[1]]
    dbfile2 <- query(ah, c("MeSHDb", "MeSH.AOR.db", "v002"))[[1]]
    dbfile3 <- query(ah, c("MeSHDb", "MeSH.PCR.db", "v002"))[[1]]
    MeSH.Hs.db <- MeSHDbi::MeSHDb(dbfileQ)
    MeSH.db <- MeSHDbi::MeSHDb(dbfile1)
    MeSH.AOR.db <- MeSHDbi::MeSHDb(dbfile2)
    MeSH.PCR.db <- MeSHDbi::MeSHDb(dbfile3)

    # ORA analysis - defining the relevant parameters
    datameshR <- read.csv("assignment_meshr.csv")
    meshParams <- new("MeSHHyperGParams",
                      geneIds = datameshR$Query,
                      universeGeneIds = datameshR$Background,
                      annotation = "MeSH.Hs.db",
                      meshdb = "MeSH.db", category = "C",
                      database = "gene2pubmed",
                      pvalueCutoff = 1e-6, pAdjust = "BH")

    # ORA analysis - Running the test
    meshR <- meshHyperGTest(meshParams)
    summary(meshR)


# QUESTION 10

  # Install “devtools” package.
    install.packages("devtools")
  # Install phexpo package from github
    devtools::install_github("GHLCLab/phexpo")
  # Load phexpo package
    library(phexpo)
    # run test
      pent <- perfFishTestChemSingle("3,4,5,3',4'-pentachlorobiphenyl",
                                     enrich_1S= TRUE)
    # Filter results to only those with bonf corrected p-value less than or equal to
0.05
      pent_filtered <- pent[pent$bonf <= 0.05,]
    # print
      print(pent_filtered)
```

```
# write results to CSV file
  write.csv(pent, file = "pentachlorobiphenyl_results.csv",
            row.names = FALSE)
```