

CoDiet

COMBATting DIET RELATED NON-COMMUNICABLE DISEASE THROUGH
ENHANCED SURVEILLANCE

D1.2 Release of an automatically annotated and manually verified corpus of 1,000 Open Access articles

Deliverable number D1.2

Work Package WP1	Technology-assisted literature triage
Task 1.1	Automated literature review of the risk of unhealthy diets on a wide spectrum of cardiometabolic diseases
Task Leader	ICL
Prepared by	Joram Posma (ICL), Antoine Lain (ICL), Tim Beck (UNOTT)
Contributors	Joram Posma (ICL), Antoine Lain (ICL), Tim Beck (UNOTT)
Version	1.1
Delivery Date	14/10/2024

Foreword

The work described in this report was developed under the project **CoDiet - Combatting Diet related non-communicable disease through enhanced surveillance** (Grant Agreement number: 101084642; Call: HORIZON-CL6-2022-FARM2FORK-01; Topic: HORIZON-CL6-2022-FARM2FORK-01-10). Any additional information, if needed, should be required to:

Project Coordinator:

Itziar Tueros – itueros@azti.es | AZTI |


WP1 Leader:

Joram Posma – j.posma11@imperial.ac.uk | ICL |

Task Leader:

Joram Posma – j.posma11@imperial.ac.uk | ICL |

Dissemination Level		
PU	Public, fully open	X
SEN	Sensitive, limited under the conditions of the Grant Agreement	
Classified R-UE/EU-R	EU RESTRICTED under the Commission Decision No2015/444	
Classified C-UE/EU-C	EU CONFIDENTIAL under the Commission Decision No2015/444	
Classified S-UE/EU-S	EU SECRET under the Commission Decision No2015/444	



Funded by the
European Union

CoDiet is part of the Horizon Europe programme supported by the European Union.

The information and views set out in this deliverable are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Document history

- (Version 1.0, 05/02/2024) – Initial version of Automated literature review of the risk of unhealthy diets on a wide spectrum of cardiometabolic diseases.
- (Version 1.1, 14/10/2024) – Revised version according to corrections required in suspension of payment letter.

TABLE OF CONTENTS

Context of deliverable within CoDiet.....	4
Information retrieval to create an Open Access corpus	4
Algorithms for automated annotation of relevant terms.....	5
Silver-standard annotated corpus of 1,000 articles.....	8
Human verification of machine annotations	11
Gold-standard annotated corpus.....	13
Output in context.....	14
Summary	17

Context of deliverable within CoDiet

Work package (WP) 1 focusses on AI-assisted literature review to support other WPs to identify data, cohorts, and links between biomedical entities to be evaluated within the CoDiet project. WP1 feeds into WP2 to provide a list of potential target biomarkers, into WP3 and WP4 to identify datasets that can be used to learn feature embeddings applicable to the new data that will come out of WP2. Furthermore, WP1 will support WP5 in creating a knowledge graph of relations between entities in the wider context of diet and non-communicable disease.

During the kick-off meeting (KOM) Task 1.0 was completed to reach a consortium-wide consensus on the scope of the literature review, specifically on data entities and phenotypes of interest. The result was to focus on metabolic syndrome and its components with multi-modal data relating to various omics and wearable technologies.

Task 1.1 involved an “Automated literature review of the risk of unhealthy diets on a wide spectrum of cardiometabolic diseases”, with a deliverable to “Release [of] an automatically annotated and manually verified corpus of 1,000 Open Access articles” to be used for optimising existing and creating new algorithms to be used for Task1.2 (“Literature review to find evidence for key physiological processes involved in the diet-related risk for cardiometabolic diseases”).

Information retrieval to create an Open Access corpus

In brief, 6 categories of search terms were used, 5 lists of inclusion terms (disease/phenotypes, diet, data, methodology, study type) and 1 list of exclusion terms, that were distilled from the consensus from the KOM. In total, 121 phenotype terms were searched in titles or abstracts (e.g. cardiometabolic syndrome, dyslipidemia, glucose response), 153 diet-related terms (e.g. caloric restriction, diet diaries, nutritional behaviour), 23 data types (e.g. image, urine, stool), 90 methodologies (amplicon sequencing, camera technology, polygenic risk score), 81 study types (e.g. cohort study, randomized controlled clinical trial, personalised nutrition), and 49 exclusion terms (cancer, NAFLD, saliva). These terms were used to create search queries for Web of Science (WoS), PubMed and PubMed Central (PMC) and from each of these, lists of identifiers (WoS ID, PMID, PMCID, DOI) were extracted.

On 7 June 2023 the search queries were run and a total of 3,372 identifiers from PubMed, 1,645 from PMC and 12,687 from WoS were found as potentially relevant for CoDiet. We used application programming interfaces (APIs) to resolve the PMC and WoS identifiers to PubMed identifiers (PMIDs), resulting in 12,112 PMIDs. We then used the open-sourced Cadmus system, developed in Python, to extract all full-text data for these articles. Cadmus serves as a solution to generate biomedical text corpora from full-text published literature. The challenge of acquiring such datasets has long hindered methodological advancements in biomedical natural language processing (bioNLP) and limited the capacity to extract knowledge from the biomedical published literature. Cadmus is highly adaptable and designed to retrieve both Open Access (OA) articles and those from publishers accessible by users via their host institutions’ library licenses. Cadmus can process documents of diverse formats, standardising their extracted content into machine-readable plain text, and organising article metadata. For CoDiet, Cadmus retrieved 10,173 publications out of 12,112 (83.99%) directly. An additional 251 publications were identified from PMC and 740 from the doi.org API. This brings the CoDiet corpus to a total of 11,164 publications (92.17% retrieval rate).

We divided the corpus into three subsets: OA publications that can be freely shared and reused, OA publications that prohibit redistribution and non-OA publications. The first set of data comprises 3,688

full-text publications which can be freely shared both within the CoDiet consortium and publicly. Our initial focus has been to apply and develop our NLP tools on this corpus as the output can be redistributed and reused in future, whereas the redistribution-prohibiting OA and non-OA articles cannot be shared in full and from these only parts of the full text surrounding text-mined annotations can potentially be shared within the legal confines of text mining. The 3,688 articles were then processed into a machine-readable format using Auto-CORPus. Auto-CORPus is a software tool designed to standardise and convert full-text research publications, particularly in the biomedical field, into machine-readable formats. It focuses on extracting text and tables from HTML and XML documents and ensures consistency in their structure and representation. Auto-CORPus converts the publications obtained from Cadmus to BioC JSON format (Figure 1) designed for exchanging text data and annotations in the life sciences domain. It aims to facilitate interoperability between various text mining tools and resources, overcoming the challenge of incompatible data formats.

```
{
  "source": "Auto-CORPus (XML)",
  "date": "20231112",
  "key": "autocorpus_fulltext.key",
  "infons": {
    "pmid": "PMC8310931",
    "doi": "10.3389/fnut.2021.691401",
    "link": "https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8310931/",
    "journal": "Frontiers in Nutrition",
    "pub_type": "Nutrition",
    "year": "2021",
    "license": "This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms."
  },
  "documents": [
    {
      "id": "PMC8310931",
      "infons": {
        "passages": [
          {
            "offset": 0,
            "infons": {
              "section_title_1": "document title",
              "iao_name_0": "document title",
              "iao_id_0": "IAO:0000305"
            },
            "text": "Fatty Acid Esters of Hydroxy Fatty Acids (FAHFAs) Are Associated With Diet, BMI, and Age",
            "sentences": [],
            "annotations": [],
            "relations": []
          },
          {
            "offset": 85,
            "infons": {
              "section_title_1": "keywords",
              "iao_name_0": "keywords section",
              "iao_id_0": "IAO:0000630"
            },
            "text": "body weight, diet, diabetes mellitus, bariatric surgery, obesity, fatty acid esters of hydroxy fatty acids",
            "sentences": [],
            "annotations": [],
            "relations": []
          },
          {
            "offset": 156,
            "infons": {
              "section_title_1": "Abstract",
              "iao_name_0": "textual abstract section",
              "iao_id_0": "IAO:0000315"
            },
            "text": "Background: Fatty acid esters of hydroxy fatty acids (FAHFAs) are a group of fatty acids with potential anti-inflammatory and anti-diabetic effects. The blood levels of FAHFAs and their regulation in humans have hardly been studied.Objective: We aimed to investigate serum FAHFA levels in well-characterized human cohorts, to evaluate associations with age, sex, BMI, weight loss, diabetic status, and diet.Methods: We analyzed levels of stearic-acid-9-hydroxy-stearic-acid (9-SAHSAs), oleic-acid-9-hydroxy-stearic-acid (9-OAHSAs) and palmitic-acid-9-hydroxy-palmitic-acid (9-PAHFA) as well as different palmitic acid-hydroxy-stearic-acids (PAHSAs) by HPLC-MS/MS with the use of an internal standard in various cohorts: A cohort of different age groups (18-25y; 40-45y; 75-85y; n = 60); severely obese patients undergoing bariatric surgery and non-obese controls (n = 36); obese patients with and without diabetes (n = 20); vegetarians/vegans (n = 10) and omnivores (n = 9); and young men before and after acute overfeeding with saturated fatty acids (SFA) (n = 15).Results: Omnivores had substantially higher FAHFA levels than vegetarians/vegans [median (25th percentile; 75th percentile) tFAHFAs = 12.82 (7.57; 14.86) vs. 5.86 (5.10; 6.71) nmol/L; P < 0.05]. Dietary overfeeding by supplementation of SFAs caused a significant increase within 1 week [median tFAHFAs = 4.31 (3.31; 5.27) vs. 6.96 (6.50; 7.76) nmol/L; P < 0.001]. Moreover, obese patients had lower FAHFA levels than non-obese controls [median tFAHFAs = 3.24 (2.30; 4.30) vs. 5.22 (4.18; 7.46) nmol/L; P < 0.01] and surgery-induced weight loss increased 9-OAHSAs level while other FAHFAs were not affected. Furthermore, significant differences in some FAHFA levels were found between adolescents and adults or elderly, while no differences between sexes and between diabetic and non-diabetic individuals were detected.Conclusions: FAHFA serum levels are strongly affected by high SFA intake and reduced in severe obesity. Age also may influence FAHFA levels, whereas there was no detectable relation with sex and diabetic status. The physiological role of FAHFAs in humans remains to be better elucidated.Trial Registration: All studies referring to these analyses were registered in the German Clinical Trial Register (https://www.drks.de/drks\_web/) with the numbers DRKS00009008, DRKS00010133, DRKS00006211, and DRKS00009797.",
            "sentences": [],
            "annotations": [],
            "relations": []
          }
        ]
      }
    }
  ]
}
```

FIGURE 1. STRUCTURE OF BIOC-STANDARD JSON FILE OF AN EXAMPLE ARTICLE (PMC8310931).

Algorithms for automated annotation of relevant terms

From Task 1.0 emerged 13 categories of potentially relevant entities to annotate, these are listed in Table 1 with a brief description. Each category was colour-coded, ensuring that overlapping categories of entities (e.g. gene and protein, and diet and food) have colours that are easily distinguishable by humans (including those that are colourblind). Some entities can overlap between categories and their allocation depends on the context (e.g. glucose intake (foodRelated) vs blood glucose (metabolite)).

TABLE 1. SUMMARY OF THE 13 CATEGORIES TO WHICH ANNOTATED ENTITIES ARE ASSIGNED, INCLUDING A BRIEF DESCRIPTION OF WHAT THEY REPRESENT.

Category name	Description
dietMethod	<i>A broad description of the type of methodology used to capture dietary/nutrition data</i>
foodRelated	<i>Food, nutrients, and diets that are mentioned in the context of intake, biomarkers measured in biofluids should not be included</i>
metabolite	<i>Small molecules measured in biofluids and tissue, includes both metabolites and lipids</i>
microbiome	<i>Microbial entities from (super)kingdoms archaea, bacteria and fungi, this can be of any taxonomic rank</i>
proteinEnzyme	<i>Protein and enzyme mentions including protein hormones and lipoproteins, this is distinct from the gene mentions although these can overlap depending on context</i>
geneSNP	<i>Gene and single-nucleotide polymorphism mentions, this can include both abbreviations for genes as well as full names</i>
diseasePhenotype	<i>Any term related to a type of disease/phenotype/medical condition that is being studied or for which the results are relevant, we are primarily interested in entities around metabolic syndrome, however any non-communicable disease entity is also of interest</i>
sampleType	<i>The sample types from which data is collected</i>
dataType	<i>The type of data or technology used by the study to analyse/measure the samples</i>
methodology	<i>A broad description of the type of general study methodology used, this is different from the diet methodology</i>
modelOrganism	<i>Descriptions of the type of organism the study is conducted with, i.e. human or animal data, this should not include microbiota</i>
populationCharacteristic	<i>A broad range of descriptors from the cohorts or populations that are studied</i>
computational	<i>Computational/statistical methods used to analyse the data</i>

To annotate the articles, a variety of methods were used (Table 2). We used dictionaries of terms that were contributed by CoDiet collaborators during the KOM, these dictionaries were supplemented with their synonyms and plural/singular forms (where relevant). In addition, we used several manually curated ontologies to search for these terms in the full-text documents. For domains where entities take the form of regular patterns, such as for single nucleotide polymorphisms (SNPs), we used regular expression (RegEx) methods to look for these patterns in the text. Finally, we used several deep learning-based methods for named-entity recognition (NER) that are both publicly available and developed in-house (and are in submission).

TABLE 2. SUMMARY OF METHODS USED TO ANNOTATE ENTITIES FROM THE 13 CATEGORIES.

Category name	Dictionaries, ontologies and algorithms used for annotation

dietMethod	CoDiet dictionary (105 terms), OBO:one (ontology)
foodRelated	CoDiet dictionary (44 terms), OBO:cdno (ontology) , OBO:foodon (ontology), UMLS:food (ontology)
metabolite	Dictionary-based annotation and normalisation (1,934,277 terms, based on HMDB and LIPID MAPS ontologies), MeLiNER (deep learning model, manuscript in preparation), TABoLiSTM (deep learning model, public)
microbiome	Dictionary-based annotation and normalisation (124,706 terms), microBERT (deep learning model, manuscript in preparation)
proteinEnzyme	BERN2 (deep learning model, public), eNzymER (dictionary-based annotation and deep learning model, public)
geneSNP	BERN2 (deep learning model, public), regular expression (for SNPs)
diseasePhenotype	CoDiet dictionary (228 terms), BERN2 (deep learning model, public), PhenoBERT (deep learning model, public)
sampleType	CoDiet dictionary (24 terms), UMLS:bodysubstance (ontology)
dataType	CoDiet dictionary (92 terms), UMLS:laboratoryprocedure (ontology)
methodology	CoDiet dictionary (136 terms), UMLS:researchactivity (ontology)
modelOrganism	CoDiet dictionary (99 terms), BERN2 (deep learning model, public)
populationCharacteristic	CoDiet dictionary (39 terms)
computational	CoDiet dictionary (44 terms), OBO:stato (ontology), OBO:obcs (ontology)

Dictionary matching involves identifying entities by comparing words in the text against a predefined list or dictionary of known entities. RegEx offer a powerful method for pattern-based entity extraction, allowing for the identification of entities based on specific character sequences or formats. Ontologies provide a structured representation of knowledge, defining relationships and hierarchies between entities, enhancing the accuracy and context of NER by incorporating semantic information as well as assigning standardised database identifiers allowing multiple synonyms of the same entity to be mapped to the same ID (named entity normalisation, NEN). Deep learning methods for NER and NEN utilise neural network architectures, such as recurrent or transformer models, to automatically learn complex patterns and dependencies in text, enabling more nuanced and context-aware entity recognition and normalisation.

We developed a hierarchical system to decide on which annotation to use in case of disagreements between methods. In this system, we first aggregate annotations with identical spans and biomedical types. This involves consolidating identifiers and methods to maintain a single, aggregated annotation with comprehensive metadata. In cases where multiple annotations share the same span but differ in biomedical types, a rule-based approach is implemented. The rule-based method follows a priority order, favouring dictionary methods populated with terms confidently identified with their biomedical types. Subsequently, deep learning methods are considered in a specific sequence (Phenobert, microBERT, MetaboLipidBERT, TABoLiSTM, eNzymER, BERN2). After this, dictionary matching takes precedence over ontology matching (curated OBO ontologies), with the Metamap ontology

considered last due to large numbers of potential false positives being annotated. When an annotation is encompassed by another (i.e. one is a complete subset of another), we retain the longest annotation. For overlapping annotations with distinct biomedical types, the rule-based decision approach is again invoked. However, if overlapping annotations share the same entity type (category), then they are merged to form a more extensive annotation that encapsulates all relevant information. This structured approach ensures a systematic resolution of discrepancies, combining the strengths of various methods while maintaining consistency and accuracy.

After performing NER across the 13 pre-defined categories, we normalised all identified entities by assigning a common identifier to similar entities, such as synonyms (e.g., “high blood pressure” and “hypertension”). Our normalisation process is based on the methods used during the NER stage. Entities identified with dictionary, ontology, and regular expression methods are normalised based on textual similarity. We used fuzzy matching to assign the identifier of the closest term from our database, accounting for differences using Levenshtein distance, which considers the number of inserted, deleted, or substituted characters. For entities identified via deep learning models, we employed contextual similarity. These deep learning models, trained on large corpora, learn relationships between words by examining their co-occurrence within sentences. Through word embeddings, extracted from our newly trained models, where words are projected into a higher-dimensional space, the model can capture semantic similarities. By computing the distance between word vectors using the cosine similarity, we identify the closest known term from our curated database (e.g. for metabolites we created a combined Human Metabolome DataBase (HMDB) and LIPID MAPS dictionary of identifiers with names/synonyms) and assign its identifier to the newly identified entity.

Silver-standard annotated corpus of 1,000 articles

The algorithms described in Table 2 were applied to the Open Access documents potentially relevant to CoDiet based on the search strategy. Across the 1,000 articles, there were 485,653 annotations in total, on average this equates to ~486 per article. Table 3 describes the number of articles that have at least one annotation from a category, and across all articles the total number of annotations per article, the total number of unique annotations, and the total number of unique identifiers (e.g. systolic blood pressure and SBP are two unique annotations that map to one identifier).

TABLE 3. SIZE OF THE SILVER-STANDARD ANNOTATED CODIET CORPUS (N=1,000).

Category name	Number of articles with at least one annotation	Number of annotations across articles		
		Total	Total unique (trigger)	Total unique identifiers
dietMethod	971	20,006	138	74
foodRelated	1,000	73,259	3,098	2,153
metabolite	981	54,465	5,712	5,446
microbiome	440	16,637	1,404	908
proteinEnzyme	690	3,372	536	352
geneSNP	943	33,398	9,444	3,948

diseasePhenotype	1,000	77,039	17,029	4,396
sampleType	999	37,835	995	526
dataType	1,000	33,142	2,740	1,296
methodology	1,000	32,147	2,679	1,119
modelOrganism	1,000	75,348	3,538	752
populationCharacteristic	947	23,279	67	32
computational	833	5,726	149	96
total	1,000	485,653	47,529	21,098

```

"documents": [
  {
    "id": "PMC8310931",
    "infons": {},
    "passages": [
      {
        "offset": 0,
        "infons": {
          "section_title_1": "document title",
          "iao_name_0": "document title",
          "iao_id_0": "IAO:0000305"
        },
        "text": "Fatty Acid Esters of Hydroxy Fatty Acids (FAHFAs) Are Associated With Diet, BMI, and Age",
        "sentences": [],
        "annotations": [
          {
            "id": "3",
            "infons": {
              "type": "metabolites",
              "identifier": "LMFA03020022",
              "annotator": "MetaboLipidBERT@codiet.eu",
              "updated_at": "2023-11-12T12:10:16Z"
            },
            "text": "Hydroxy",
            "locations": [
              {
                "offset": 21,
                "length": 7
              }
            ]
          },
          {
            "id": "1",
            "infons": {
              "type": "dietMethod",
              "identifier": "CoDiet_B2_4",
              "annotator": "dictionary@codiet.eu",
              "updated_at": "2023-11-12T12:10:16Z"
            },
            "text": "Diet",
            "locations": [
              {
                "offset": 70,
                "length": 4
              }
            ]
          },
          {
            "id": "2",
            "infons": {
              "type": "populationCharacteristic",
              "identifier": "CoDiet_G1_1",
              "annotator": "dictionary@codiet.eu",
              "updated_at": "2023-11-12T12:10:16Z"
            },
            "text": "Age",
            "locations": [
              {
                "offset": 85,
                "length": 3
              }
            ]
          }
        ],
        "relations": []
      }
    ]
  }
]

```

FIGURE 2. STRUCTURE OF BIOC-STANDARD JSON FILE OF A MACHINE-STANDARD ANNOTATED ARTICLE (PMC8310931).

The documents are created in a BioC-compliant manner, where each annotation is marked with a unique identifier, the text, location and span, the type, an external identifier (ontology), the annotator (algorithm) and date of annotation, see Figure 2. The same file structure is used for the gold-standard corpus as input to TeamTat (as BioC-compliant XML file).

Human verification of machine annotations

We recruited annotators from within the CoDiet consortium via an open call for recruitment during a consortium-wide meeting in October 2023 and via email. Additionally, we recruited further volunteers from within one of the institutions. A total of 44 people initially volunteered from five partners (ICL, AUTH, AZTI, CICBIO, UVEG). We ran a dedicated online meeting for all potential annotators to demonstrate the software used (TeamTat, see below), and to run a demo annotation task with all people attending the meeting (28). During this meeting the participants were briefed on the categories types (Table 1), were given examples for each of these 13 categories (Table 4) and we introduced 2 additional categories to help annotators (Table 5) in cases where they are unsure of the category and/or have identified potentially relevant entity types not attributable to any of the other categories. While the silver-standard corpus contains identifiers for each entity, the task is focussing on NER and not normalisation (NEN) of these entities to common identifiers. Thus, the annotators were not asked to add identifiers for the annotations they make. The meeting was recorded for the other volunteers and shared with them after the session.

TABLE 4. EXAMPLES OF ENTITIES MATCHING TO EACH CATEGORY THAT WERE GIVEN TO ANNOTATORS AS PART OF THE TASK BRIEFING.

Category name	Example 1	Example 2	Example 3	Example 4	Example 5
dietMethod	food frequency questionnaire	nutritional epidemiology	dietary recommendations	personalised nutrition	dietary assessment
foodRelated	Mediterranean diet	dietary fibre	high-fat diet	omega-3	apple
metabolite	sphingomyelin (d18:0/17:0)	Trimethylamine N-oxide	cholesterol	PC(34:0)	glucose
microbiome	<i>Methanobrevibacter smithii</i>	<i>Akkermansia</i> spp.	<i>Candida albicans</i>	<i>Firmicutes</i>	<i>E. coli</i>
proteinEnzyme	alanine aminotransferase	LDL-cholesterol	oxidoreductase	glucagon	CRP
geneSNP	lipoprotein lipase gene	rs1571960363	CRP gene	MALAT1	GNPAT
diseasePhenotype	insulin resistance	atherosclerosis	prehypertension	obesity	stroke
sampleType	24hr urine collections	fasted	stool	video	gut
dataType	16s rRNA gene sequencing	micro-camera assessment	bioimpedance	lipidomic	HPLC-MS
methodology	randomised controlled clinical trial	public health policy	personal monitoring	free-living cohort	cross sectional
modelOrganism	<i>Rattus norvegicus</i>	<i>C. elegans</i>	patient	human	mouse
populationCharacteristic	physical activity	family history	lifestyle	female	age
computational	logistic regression	machine learning	paired t-test	correlation	AI model

TeamTat is a collaborative, online text annotation tool, designed to facilitate the annotation process for large document collections to ready these for downstream tasks such as training NLP models. It has a user-friendly interface for both project managers and annotators enabling efficient labelling. Anonymity features are available to reduce bias, while quality control is performed after annotations are done to ensure data accuracy. In our implementation of TeamTat, we added a customisable visibility tab to help annotators focus on specific categories they have expertise for. Jointly with the ICL Research Software Engineering teams, we hosted a local installation of the TeamTat software on an Azure cloud platform to allow (credentialised) access for any CoDiet member (<https://teamtat-prod-app.azurewebsites.net/>).

After the demo meeting we reached out to all volunteers to ask them for their self-perceived expertise for the 13 categories, their time commitment (number of articles per week/in total) and assigned articles in batches to the 38 volunteers that responded. Batches were released weekly and assigned semi-randomly, i.e. annotators were assigned articles that had entities relevant to their expertise (e.g. Table 6) as not all articles had all categories but were then assigned their specified number of articles for each batch at random. Three annotators that initially verified their involvement dropped out after having been assigned articles. Once an annotation round is started, TeamTat does not allow re-assigning articles to annotators (based on their user-id), hence we updated the login details of each user-id to give other annotators a secondary login to complete these articles. Over the course of the first 6 weeks of the task, we scheduled 1-2 weekly drop-in sessions via MS Teams that any annotator can join and share their screen to get help/advice. These sessions were initially attended by 15-20 annotators each time, but over time (as instructions were clarified and annotators gained experience) this was reduced to 1-2 individuals. As a result of these drop-in sessions we updated the materials on the instructions and added further examples to a [live Google Doc FAQ](#) to help the annotators.

TABLE 5. TWO ADDITIONAL CATEGORIES INCLUDED IN TEAMTAT FOR ANNOTATORS TO USE WITH THEIR DESCRIPTIONS, PROVIDED AS PART OF THE TASK BRIEFING.

Category name	Description
unsureCoDiet	<i>A broad category to be used when a term is relevant for CoDiet but no specific category can be assigned by the annotator. This can be used to indicate any of the above categories when unsure about choosing a single one.</i>
potential	<i>Any entity the annotator believes may be relevant but not sure if it is within the scope of the project.</i>

A total of 12 batches were released until the end of January 2024, culminating in a total of 1,000 articles having been seen by annotators. To improve the quality of the annotated corpus, we assigned 2 annotators to each article, thereby reaching a total number of 500 documents having been annotated by machine (see annotation depth in Table 6) followed by 2 humans independently annotating the document and verifying the machine-annotations by the algorithms outlined in Table 2. An example of the TeamTat interface with machine-annotations can be seen in Figure 3.

TABLE 6. DESCRIPTION OF THE SUBSET OF 500 ARTICLES FROM THE SILVER-STANDARD ANNOTATED CODIET CORPUS USED FOR HUMAN ANNOTATIONS.

Category name	Number of annotations across articles
---------------	---------------------------------------

	Number of articles with at least one annotation	Total	Total unique (trigger)	Total unique identifiers
dietMethod	488	10,930	115	66
foodRelated	500	38,467	2,123	1,568
metabolite	489	26,679	3,448	2,033
microbiome	217	7,662	895	618
proteinEnzyme	342	1,587	345	239
geneSNP	475	16,733	5,083	2,339
diseasePhenotype	500	39,513	9,655	3,026
sampleType	500	18,099	611	373
dataType	500	15,725	1,773	928
methodology	500	16,261	1,790	836
modelOrganism	500	37,577	2,082	491
populationCharacteristic	482	11,687	62	31
computational	416	2,996	114	78
total	500	243,916	28,096	12,626

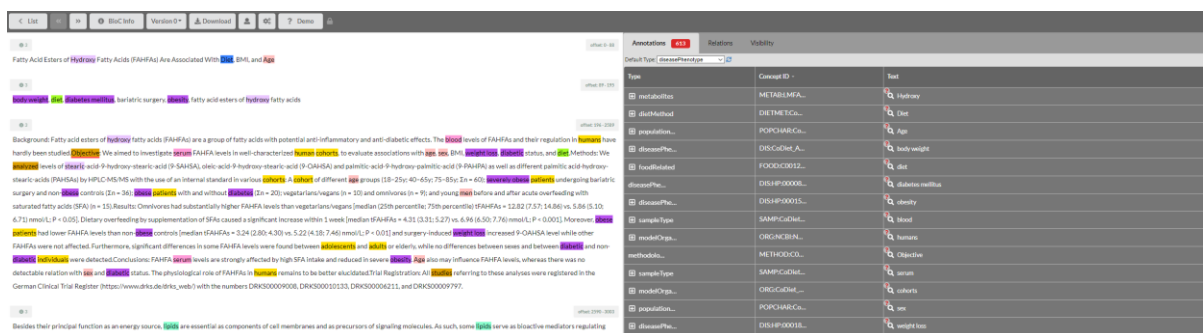
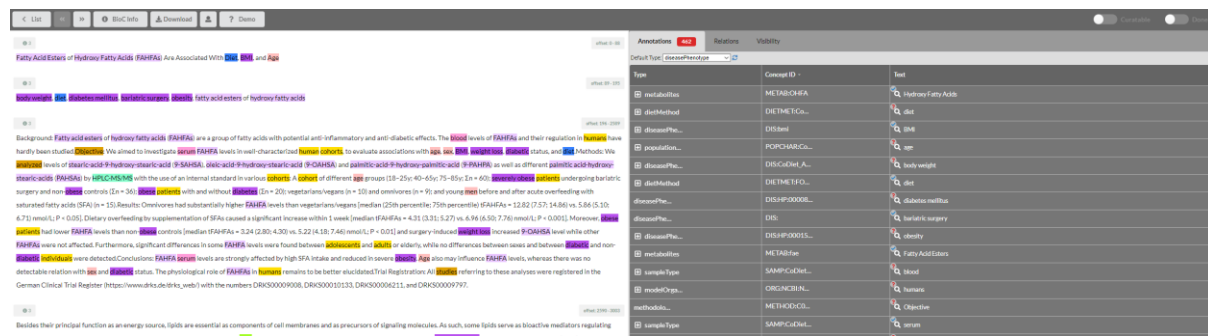


FIGURE 3. STRUCTURE OF BIOC-STANDARD XML FILE OF AN EXAMPLE ARTICLE (PMC8310931) IN TEAMTAT, MACHINE-ANNOTATIONS SHOWN AS ANNOTATORS SEE THEM.

Gold-standard annotated corpus

The annotators reviewed the full-text articles independently, i.e. pairs of annotators can only see the machine annotations but not annotations by the other annotator. This allows for the assessment of inter-annotator agreement. The human annotator-verified machine annotations are combined with new annotations made by the human annotator and visualised in the TeamTat interface (Figure 4). The right side of the window groups the annotations by category and concept ID (unique ontology

identifier) and indicates whether the annotation has been verified (blue tick) or not yet (red magnifying glass).



The screenshot shows a web interface for viewing a document with annotations. On the left, a text document is displayed with various colored highlights and icons. On the right, a table lists the annotations with columns for Type, Concept ID, and Text. The table contains the following entries:

Type	Concept ID	Text
metabolic	METABCHFA	HydroxyFattyAcids
dietMethod	DIETMETCAL	diet
diseasePhen	DISDIAB	diab
population	POPCHARGOL	pop
diseasePhen	DISCOSTA	body weight
dietMethod	DIETMETFO	diet
diseasePhen	DISDIAB00000	diabetic mellitus
diseasePhen	DIS	bariatric surgery
diseasePhen	DISDIAB00001	obesity
metabolic	METAFAT	FattyAcidEsters
sampleType	SAMPCHLAD	blood
modelDyna	ORGCHIBAL	humans
methodical	METHCHDCAL	Objective
sampleType	SAMPCHLAD	urine

FIGURE 4. STRUCTURE OF BIOC-STANDARD XML FILE OF AN EXAMPLE ARTICLE (PMC8310931) IN TEAMTAT, AFTER AN ANNOTATOR HAS ADDED ANNOTATIONS.

The silver-standard corpus had a total of 243,916 (Table 6) annotations, and the 12 batches of the human annotated corpus have a total of 275,138 by the end of the annotation task. However, annotators are able to go back to their annotations after they finish these to make further changes/additions. These are monitored as each annotation receives a unique timestamp.

Output in context

Our gold-standard corpus with 500 full-text documents has over 275K annotations in total. We surveyed publicly available bioNLP corpora that are used for training and/or testing new NLP algorithms, and note their document types, number of documents, entity types and number of annotations (Table 7). In comparison, our gold-standard corpus is the largest corpus in biomedical natural language processing in terms of total number of annotations; the next largest corpus publicly available in terms of number of entities is the BioCreative VII-ChemProt with 108K annotated proteins and chemicals across 4,250 abstracts.

Overall, over half of the available corpora focus only on abstracts, with the largest full-text corpus consisting of 499 full-text documents and a total of 88K annotations. Our full-text corpus contains almost exactly the same amount of documents but does so with more than 3x as many annotations. Only two full-text corpora contain more annotations per document than ours (~550), the CellFinder corpus of 20 articles has ~801 annotations per document and the 97 documents from the CRAFT corpus have ~1,030 annotations on average, but both are considerably smaller and more specialised.

TABLE 7. PUBLICLY AVAILABLE CORPORA WITH ANNOTATED BIOMEDICAL ENTITIES WITH RELEVANCE TO CoDIET. ENTITY CATEGORIES THAT ARE DIRECTLY RELEVANT TO CoDIET ARE HIGHLIGHTED IN BOLD, CATEGORIES THAT ARE PARTIALLY RELEVANT ARE UNDERLINED.

Corpus	Type of documents (number)	Annotated entities	Annotation level (number of annotations)
BioCreative I Gene Mention Recognition (BC1GM)	Sentences (10,000)	genes	Gold (12,000)
BioCreative II Gene Mention Recognition (BC2GM)	Sentences (20,000)	genes	Gold (44,500)
BioCreative III Interaction Method Task (BC3IMT)	Full-text articles (2,590)	genes, proteins, species	Gold (5,664)
BioCreative IV CHEMicals Disease Named Entity Recognition (ChemDNER)	PubMed abstracts (10,000)	<u>chemicals</u>	Gold (84,355)
BioCreative V Chemical Disease Relation (BC5CDR)	PubMed abstracts (1,500)	<u>chemicals</u> , diseases	Gold (13,343)
BioCreative V CHEMicals Disease Named Entity Recognition patents (ChemDNER patents)	Patent abstracts (21,000)	<u>chemicals</u> , genes	Gold (99,634)
BioCreative VI chemical-protein interactions (ChemProt)	Abstracts (1,682)	<u>chemicals</u> , proteins	Gold (42,608)
BioCreative VII chemical-protein interactions (DrugProt)	Abstracts (4,250)	<u>drugs</u> , proteins	Gold (108,387)
BioCreative VIII Genetic Phenotype Extraction	Sentences (2,170)	phenotypes	Gold (3,501)
Biomedical entity Relation ONcology Corpus (BRONCO)	Full-text articles (108)	cell lines, diseases , <u>drugs</u> , genes, variants	Gold (403)
CellFinder	Full-text articles (20)	anatomical parts, biological processes, cell components, cell lines, cell types, genes, proteins, species	Gold (16,026)
Colorado Richly Annotated Full-Text Corpus (CRAFT)	Full-text articles (97)	anatomical entities, <u>biological taxa</u> , <u>biological processes</u> , biomacromolecular entities and sequences, cellular and extracellular components and regions, cell types, <u>chemicals</u> , molecular function, chemical reactions, proteins	Gold (99,907)
Drug-Drug Interactions (DDI)	DrugBank excerpts (792), MEDLINE abstracts (233)	<u>drugs</u>	Gold (18,502)

Corpus	Type of documents (number)	Annotated entities	Annotation level (number of annotations)
Europe PMC	Full-text articles (300)	diseases, genes, proteins	Gold (72,378)
GENome Information Acquisition (GENIA)	MEDLINE abstracts (2,000)	anatomical entities, cell types, <u>chemicals</u> , DNA, organism, proteins , RNA, <u>tissue type</u>	Gold (93,293)
GeneExpression Text Miner corpus (GETM)	MEDLINE abstracts (150)	anatomical locations, genes	Gold (656)
GeNomics And Informatics (GNI)	Full-text articles (499)	cell lines, cell types, DNA, proteins , RNA	Gold (88,629)
JNLPBA	PubMed abstracts (2,000)	cell lines, cell types, DNA, proteins , RNA	Gold (59,963)
Linnaeus	Full-text articles (100)	species	Gold (4,077)
MedTag	Sentences (25,965)	domains, genes, proteins , sequences, sites	Gold (>45,000)
Multi-Level Event Extraction (MLEE)	PubMed abstracts (262)	anatomical locations, cell types, disease, drugs, genes, proteins, sample type	Gold (8,227)
NCBI-disease	PubMed abstracts (793)	disease	Gold (6,892)
NLM-Chem	Full-text articles (150)	<u>chemicals</u>	Gold (40,467)
NLM-Gene	PubMed abstracts (550)	genes, species	Gold (15,581)
OpenMinTeD	Abstracts (200), full-text articles (100)	biological activity, <u>chemicals</u> , metabolites, proteins, species	Gold (86,532)
Phenotype-Gene Relations (PGR)	PubMed abstracts (1,712)	genes, phenotypes	Silver (19,511)
Species-800	PubMed abstracts (800)	species	Gold (3,708)

Summary

Using text-mining, natural language processing and domain-expert involvement we have created the largest biomedical entity corpus known to us at this time. The corpus is currently in the process of normalisation, i.e. for all annotated entities we are adding in unique identifiers from databases and ontologies. Once this process is complete the dataset will be shared in its entirety as part of an Open Access publication and hosted in a public data repository to ensure longevity of the corpus. Adding in identifiers for each annotation will ensure more widespread reuse of the dataset, permitted due to having used only Open Access articles with CC-BY licences that permit redistribution, as it will be possible to be used by others to train new named-entity recognition and named-entity normalisation algorithms, which in turn can be implemented in the CoDiet project. Other corpora use teams of annotators that are paid for their work, here we only used volunteers, and we recommend others undertaking the annotation of new corpora in future with volunteers to offer them a suitable incentive. In our case our annotators will all be credited in the data set release, acknowledged in the arising publication and some included as co-authors. Moreover, this corpus has explored several categories for which no ontologies exist and/or for which existing ontologies were deemed too broad (and resulting in many false positive annotations) or too narrow (not capturing enough of the rich biomedical entities that exist in the CoDiet corpus). These ontologies can now be developed using the annotated dataset with clear examples for each type.