CoDiet

COMBATTING DIET RELATED NON-COMMUNICABLE DISEASE THROUGH ENHANCED SURVEILLANCE

D1.1 Stratified list of precision medicine initiatives, ranked according to their potential synergy with CoDiet

Deliverable number D1.1

Work Package WP1	Technology-assisted literature triage
Task 1.3	Monitoring of existing datasets and large cohorts in Europe
Task Leader	CICBIO
Prepared by	Joram Posma (ICL)
Contributors	Joram Posma (ICL), Nieves Embade (CICBIO), Antoine Lain (ICL)
Version	1.0
Delivery Date	10/07/2023

Funded by the European Union
UK Research and Innovation

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:
Oscar Millet – omillet@cicbiogune.es | CICBIO |

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.

TABLE OF CONTENTS

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 1.3 involved the “Monitoring of existing datasets and large cohorts in Europe”, with a deliverable of sharing a “Stratified list of precision medicine initiatives, ranked according to their potential synergy with CoDiet” (D1.1).

Data collection methodology and results

Publicly available scholarly databases, including Pubmed and Google Scholar, were searched using a variety of search queries including "large number of samples", "large human cohorts", "studies with large numbers of individuals", "large cohorts", "large populations", "precision medicine", "longitudinal studies", "biobanks", etc. Three individual researchers, at CIC bioGUNE, independently scanned titles and abstracts, specifically focussing on statements pertaining to studies containing a large number of individuals, e.g. "17,000 individuals were recruited". For each potentially relevant article, links to websites related to initiatives/data were identified and these websites were then scanned with relevant information extracted such as the initiative title, participating countries, the disease focus, and data access (public/restricted/private) (**Figure 1**). The initial focus was to identify precision medicine initiatives in Europe, however during the analysis this was expanded to worldwide initiatives.

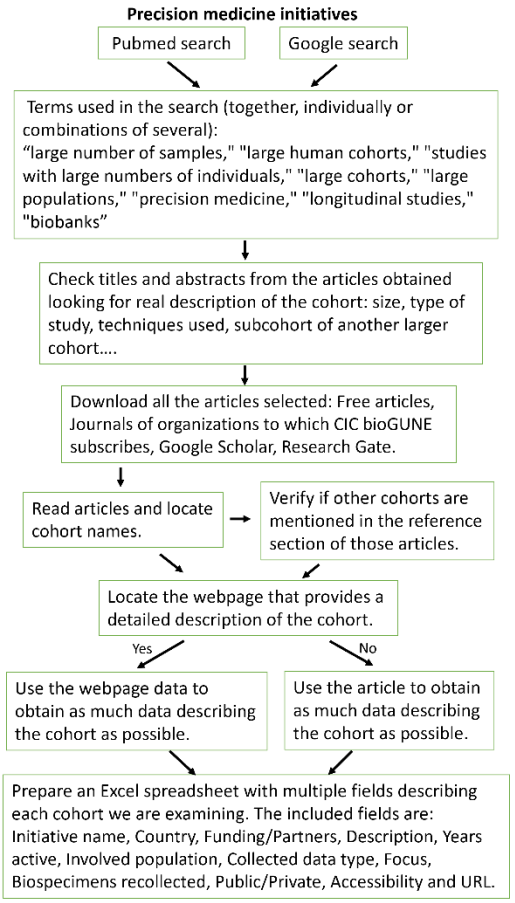

FIGURE 1. FLOW DIAGRAM OF SEARCH STRATEGY FOR IDENTIFYING PRECISION MEDICINE INITIATIVES.

Data curation and results

The 34 initiatives identified in the first step were passed on to the NLP team at Imperial College. Two researchers independently evaluated the data by web scraping (text mining) information from the web pages (using the initial URL), including delineating the geographical area each identified initiatives covered (from 'EU' or 'Asia' to individual countries/partners) (Figure 2).

The same researchers then extracted summary text, links to data, projects/studies and publications

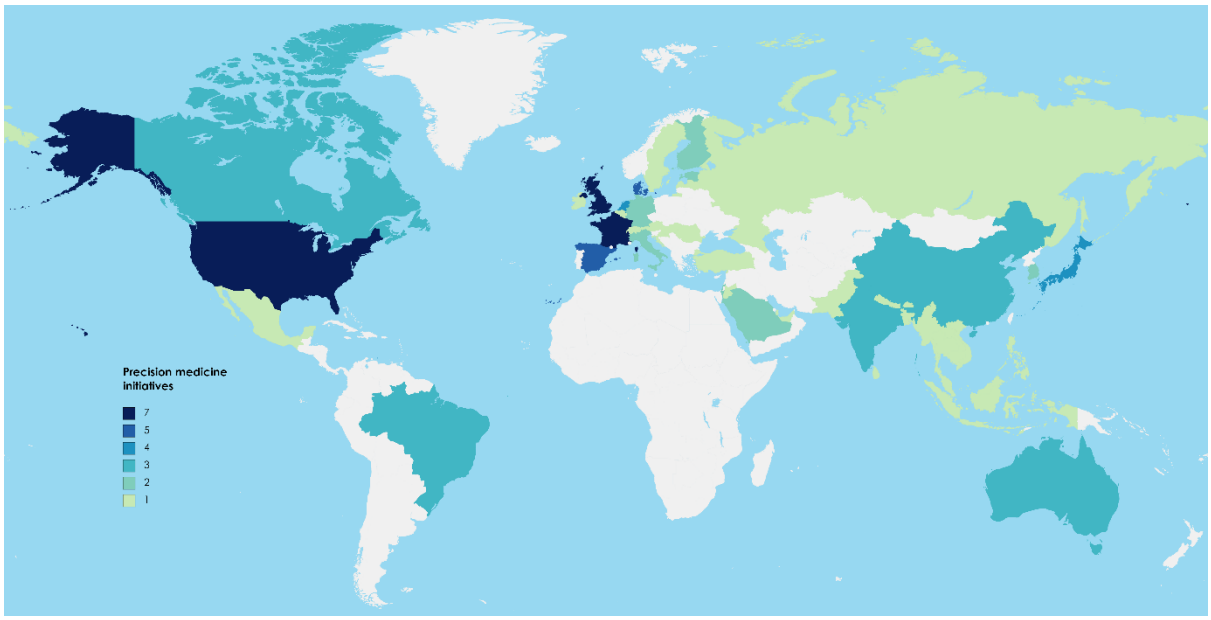

FIGURE 2. OVERVIEW OF GEOGRAPHICAL LOCATIONS LINKED TO PRECISION MEDICINE INITIATIVES. MAP CREATED USING MAPCHART.NET.

(where available) and evaluated whether the disease focus was within scope for CoDiet. Most initiatives excluded at this stage related to those that were evaluated to be specifically tailored to cancer, rare disease or other non-communicable disease phenotypes explicitly listed in titles and summary text (Figure 3).

From the 13 initiatives with text data that passed the initial screening (Figure 3) a total of 927 unique identifiers were extracted from project/publication webpages. These were comprised of 380 DOIs (Digital Object Identifiers), 271 PMIDs (PubMed Identifiers), 177 URLs (Uniform Resource Locators) and 99 PMCIDs (PubMed Central Identifiers), where these identifiers can relate to the same publication. These were then matched using various convertors and application programming interfaces (APIs), including NCBI's (National Center for Biotechnology Information) idconv and ESearch, PaperPile's DOI to RIS and CrossRef's XML API. This resulted in identifying 379 unique publications arising from these initiatives, in addition to the 177 URLs of funded projects (Figure 4), with a 100% retrieval rate of associated texts. Titles, abstracts and project descriptions were extracted through APIs (PaperPile, NCBI's EFetch), web scaping and PDF (Portable Document Format) to text convertors.

These free text files were then submitted to the same search strategy as for articles identified for D1.2 and D1.3 through the output of Task 1.0. In brief, 6 categories of search terms are used, 5 inclusion terms (phenotypes, diet, data, methodology, study type) and 1 list of exclusion terms. In total, 121 phenotype terms are 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).

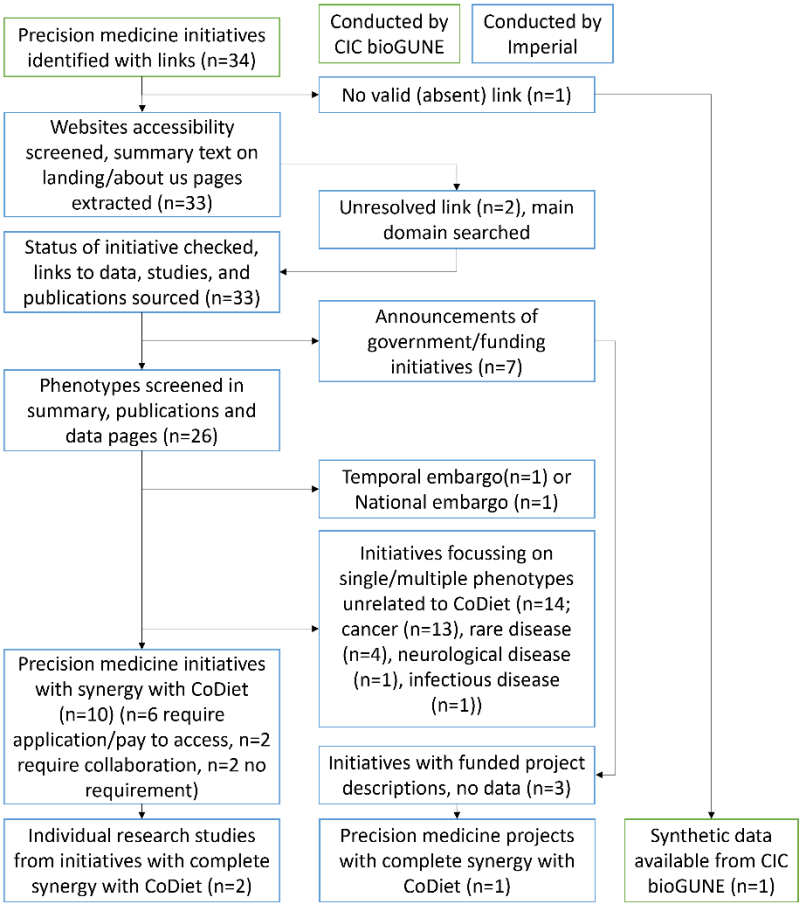

FIGURE 3. FLOW DIAGRAM OF CURATION STRATEGY OF IDENTIFIED PRECISION MEDICINE INITIATIVES AND ASSOCIATED DATA (SUMMARIES, STUDIES, PUBLICATIONS).

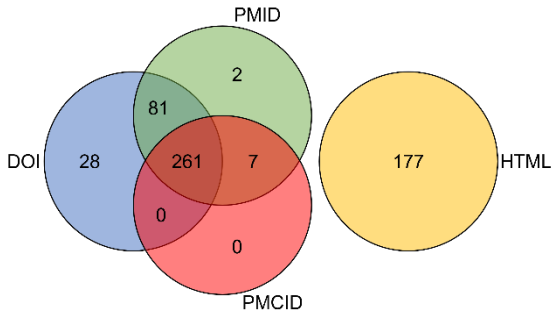

FIGURE 4. VENN DIAGRAM OF OVERLAP BETWEEN IDENTIFIERS.

Deliverable output

556 texts were then scored according to the search terms (Task 1.0) for each category separately. These were assigned a score from -1 to 5, where -1 indicates any exclusion term was matched, and

0-5 indicate how many of the inclusion term categories had 1 or more terms found, with nearly 40% matching an exclusion term and over half matching at least 1 term from any category (Table 1). For each of the texts the matched terms were extracted (see summary in Table 2), and those studies matching to 3 or more categories (n=43) are summarised in Table 3, ranked first by score and next by number of overall terms matched. The overall list of 281 studies from these initiatives has been shared with WP3 and WP4 internally, including access requirements.

TABLE 1. SUMMARY OF THE CODIET-SYNERGY SCORES ACROSS 556 TEXTS FROM 13 INITIATIVES.

Score	Number of texts (%)
5	3 (0.54%)
4	10 (1.80%)
3	30 (5.40%)
2	69 (12.41%)
1	169 (30.40%)
0	57 (10.25%)
-1	218 (39.21%)

TABLE 2. ALL TERMS IDENTIFIED FROM 556 TEXTS FOR EACH OF THE 6 CATEGORIES OF INCLUSION AND EXCLUSION TERMS RELEVANT TO CODIET.

Term category	Matched terms (number of times matched)
Phenotype	obesity (29), hypertension (18), lipids (11), blood pressure (10), obese (10), triglyceride (10), triglycerides (10), overweight (8), hypercholesterolemia (7), hyperlipidemia (7), dyslipidemia (4), insulin sensitivity (4), weight loss (4), body weight (3), high cholesterol (3), metabolic syndrome (3), waist circumference (3), body composition (2), glycemic control (2), weight gain (2), blood glucose (1), body fat distribution (1), dyslipidaemia (1), elevated blood pressure (1), fasting glucose (1), glucose metabolism (1), glycemia (1), insulin resistance (1), insulin resistant (1), oral glucose tolerance test (1), trunk fat percentage (1), waist-to-hip ratio (1)
Diet-related	diet (46), nutrition (27), nutritional (14), saturated fat (5), polyunsaturated fat (4), dietary intervention (3), food intake (3), high-fat diet (3), malnutrition (3), dietary fat (2), dietary habits (2), dietary intake (2), dietary patterns (2), personalised nutrition (2), caloric restriction (1), dietary assessment (1), dietary assessments (1), dietary data (1), dietary fiber (1), effect of diet (1), food frequency questionnaire (1), healthy eating (1), intermittent fasting (1), ketogenic diet (1), low-carbohydrate diet (1), plant-based food (1), precision nutrition (1), specific nutrients (1), vegetables (1), western diet (1)
Data	blood (59), plasma (21), gut (20), urine (13), image (10), serum (9), urinary (7), membrane (6), images (5), gastrointestinal (3), video (3), fecal (2), cell membrane (1), feces (1), imagery (1), stool (1)
Methodology	genome (129), genomic (82), sequencing (80), genomics (33), genotype (30), genomes (24), metabolites (12), microbiota (12), polygenic risk score (10), genotypes (9), wearable (9), gene expression (8), microbiome (8), epigenetic (7), metabolomic (6), metabolomics (6), gut microbiome (5), wearable device (5), next generation sequencing (3), transcriptomics (3), camera (2), high-throughput sequencing (2), metabolic profiles (2), metabolome (2), amplicon sequencing (1), genetic risk score (1), genotyping arrays (1), lipidomic (1), lipidomics (1), lipoproteins (1), omic technologies (1), shotgun sequencing (1), spectroscopy (1)

Study type	precision medicine (62), intervention (61), biomarker (53), cross-sectional (48), biomarkers (44), cohort study (30), clinical trials (16), epidemiology (13), personalized medicine (12), high-throughput (8), meta-analyses (6), personalised approach (5), case study (3), individual level (3), observational studies (3), population level (3), apps (2), assessment tool (2), biomarker panel (2), epidemiological studies (2), human intervention (2), individual variability (2), mobile technology (2), personalised nutrition (2), randomised controlled trial (2), clinical population (1), high throughput (1), human trials (1), individual response (1), individual variation (1), intervention studies (1), intervention study (1), mobile apps (1), time series (1)
Exclusion terms	cancer (88), oral (32), skin (30), stress (28), depression (21), autoimmune (20), cancers (17), anxiety (16), physical activity (13), inflammation (12), biopsy (11), asthma (9), dementia (9), saliva (8), cognitive impairment (6), hepatitis (6), allergy (4), rare diseases (4), type 1 diabetes (4), chronic obstructive pulmonary disease (3), inflammatory response (3), kidney failure (2), pneumonia (2), sarcopenia (2), colon cancer (1), fetal (1), gout (1), migraine (1)

TABLE 3. SUMMARY OF THE TOP 43 PRECISION MEDICINE INITIATIVES WITH RELEVANT DATA FOR CODIET WP3 AND WP4.

Score (missing categories)	Initiative	Identifier/link	Matched terms
5	Brazilian Initiative on Precision Medicine	doi: 10.1007/s12020-023-03356-0	body composition, obese, obesity, overweight, nutrition, plasma, fecal, gut microbiome, microbiome, cross-sectional
5	Brazilian Initiative on Precision Medicine	doi: 10.1152/ajpendo.00231.2022	obese, obesity, diet, food intake, blood, serum, metabolomics, intervention
5	Health-RI	https://www.health-holland.com/project/2022/2022/impact-wide-variety-dietary-lipids-microbiota-composition-and-functionality	lipids, dietary fat, diet, gut, gut microbiome, microbiome, microbiota, human intervention
4 (no: Phenotypes terms)	FarGen Project	doi: 10.1093/ibd/izab355 PMID: 35138361 PMCID: PMC9247847/	diet, dietary patterns, food frequency questionnaire, fecal, gut, amplicon sequencing, sequencing, genome, microbiota, cross-sectional, epidemiological studies
4 (no: Phenotypes terms)	Health-RI	https://www.health-holland.com/project/2022/2021/intestine-chip-integrated-immune-and-microbiota-compartments	diet, dietary intervention, nutrition, nutritional, personalised nutrition, specific nutrients, gut, microbiota, human trials, intervention, personalised nutrition

4 (no: Study methodology)	Japan Genomic Medicine Program	doi: 10.1073/pnas.1912573116 PMID: 31685604 PMCID: PMC6876247	weight loss, intermittent fasting, ketogenic diet, low-carbohydrate diet, diet, nutrition, plasma, gut, microbiota
4 (no: Data methodology)	Health-RI	https://www.health-holland.com/project/2022/2018/can-seaweed-reduce-blood-glucose-obese-type-2-diabetes-patients	body weight, obese, overweight, blood glucose, diet, dietary intervention, blood, plasma, intervention
4 (no: Diet terms)	MyCode Community Health Initiative	doi: 10.1001/jama.2017.0972 PMID: 28267856 PMCID: PMC5664181	lipids, triglyceride, triglycerides, plasma, sequencing, genotype, lipoproteins, cross-sectional
4 (no: Data methodology)	Health-RI	https://www.health-holland.com/project/2017/nutrition-for-an-improved-muscle-blood-flow-and-insulin-sensitivity	insulin resistance, insulin resistant, insulin sensitivity, diet, nutrition, nutritional, blood, intervention
4 (no: Study methodology)	Health-RI	https://www.health-holland.com/project/2020/2015/understanding-biological-effects-n-3-fatty-acids-different-lipid-sources-define	lipids, triglyceride, triglycerides, polyunsaturated fat, saturated fat, diet, gut, microbiome
4 (no: Phenotypes terms)	Health-RI	https://www.health-holland.com/project/2022/2021/non-invasive-continuous-gut-microbial-fermentation-measurement-health-and-disease	diet, gut, gut microbiome, metabolites, microbiome, microbiota, intervention
4 (no: Study methodology)	Japan Genomic Medicine Program	doi: 10.1126/science.aaw8429	metabolic syndrome, diet, dietary habits, gut, microbiota
4 (no: Data types)	ERA PerMed	https://erapermed.isciii.es/wp-content/uploads/2021/01/Newsletter-ERA-PerMed_final.pdf#9	metabolic syndrome, diet, genome, microbiome, intervention
3 (no: Phenotypes terms, Diet terms)	MyCode Community Health Initiative	doi: 10.1126/science.aaf6814 PMID: 28008009	blood, high-throughput sequencing, sequencing, genomic, genomics, high-throughput, precision medicine
3 (no: Phenotypes terms, Study methodology)	Brazilian Initiative on Precision Medicine	doi: 10.1186/s40168-023-01520-2 PMID: 37101209 PMCID: PMC10131329	diet, gut, video, gut microbiome, microbiome, microbiota, transcriptomics
3 (no: Data methodology, Study methodology)	Brazilian Initiative on Precision Medicine	doi: 10.3390/ijms24021729 PMID: 36675244 PMCID: PMC9861800	insulin sensitivity, obese, obesity, weight loss, high-fat diet, diet, plasma
3 (no: Data methodology,	Health-RI	https://www.health-holland.com/project/2023/2022/it-takes-guts-bbb	triglyceride, triglycerides, polyunsaturated fat,

Study methodology)			saturated fat, diet, plasma, gut
3 (no: Phenotypes terms, Diet terms)	Swiss Personalized Health Network	https://sphn.ch/wp-content/uploads/2019/11/2018DRI01_Probst-Hensch_Lay_summary_20190306.pdf	urine, blood, image, images, epigenetic, biomarker, biomarkers
3 (no: Phenotypes terms, Diet terms)	Swiss Personalized Health Network	https://sphn.ch/wp-content/uploads/2023/03/DEM-2022-01_Lay-Summary.pdf	plasma, genomic, genomics, metabolites, metabolomic, metabolomics, clinical trials
3 (no: Diet terms, Study methodology)	Precision Medicine Initiative All of Us	doi: 10.1161/circgen.122.003946 PMID: 36334310 PMCID: PMC9812363	blood pressure, elevated blood pressure, blood, sequencing, genome, polygenic risk score
3 (no: Phenotypes terms, Diet terms)	Precision Medicine Initiative All of Us	doi: 10.1016/j.xkme.2022.100582 PMID: 36712313 PMCID: PMC9879977	blood, genome, genomic, genomics, cohort study, observational studies
3 (no: Diet terms, Data types)	Precision Medicine Initiative All of Us	doi: 10.1038/s41467-021-27751-1 PMID: 35013250 PMCID: PMC8748496	hyperlipidemia, hypertension, gene expression, biomarker, biomarkers, high-throughput
3 (no: Diet terms, Study methodology)	MyCode Community Health Initiative	doi: 10.1161/circresaha.117.311145 PMID: 28506971 PMCID: PMC5523940	lipids, triglyceride, triglycerides, plasma, sequencing, genotype
3 (no: Data types, Data methodology)	Health-RI	https://www.health-holland.com/project/2015/liver-fat-insulin-sensitivity-and-diabetes-and-cardiovascular-risk	insulin sensitivity, obesity, lipids, dietary fat, diet, intervention
3 (no: Phenotypes terms, Data methodology)	Health-RI	https://www.health-holland.com/project/2018/probiotics-and-vitamin-k2-status-in-healthy-and-diseased-people	vegetables, diet, gut, human intervention, intervention, intervention studies
3 (no: Phenotypes terms, Data methodology)	Health-RI	https://www.health-holland.com/project/2021/2020/detect-halt-and-repair-early-lung-damage	diet, blood, intervention, biomarker, biomarkers, precision medicine
3 (no: Phenotypes terms, Diet terms)	Precision Medicine Initiative All of Us	doi: 10.1007/s00125-023-05912-9 PMID: 37148359 PMCID: PMC10244266	blood, genome, genomes, genotype, precision medicine
3 (no: Phenotypes terms, Study methodology)	Japan Genomic Medicine Program	doi: 10.1371/journal.pbio.3000813 PMID: 32991574 PMCID: PMC7524008	dietary fiber, diet, gastrointestinal, gut, microbiota
3 (no: Phenotypes	Health-RI	https://www.health-holland.com/project/2022/2022/	diet, nutrition, gut, gut microbiome, microbiome

terms, Study methodology)		healthier-guts-through-prebiotics-made-fungal-enzymes	
3 (no: Phenotypes terms, Data types)	ERA PerMed	https://erapermed.isciii.es/wp-content/uploads/2023/02/Newsletter-January-23_final1.pdf#16	diet, dietary patterns, metabolites, biomarker, biomarkers
3 (no: Phenotypes terms, Diet terms)	ERA PerMed	https://erapermed.isciii.es/wp-content/uploads/2023/02/Newsletter-January-23_final1.pdf#24	urinary, plasma, sequencing, genome, high throughput
3 (no: Phenotypes terms, Diet terms)	ERA PerMed	https://erapermed.isciii.es/wp-content/uploads/2022/01/Newsletter-January-22-final_comp.pdf#22	blood, genomic, genomics, biomarker, biomarkers
3 (no: Diet terms, Data methodology)	ERA PerMed	https://erapermed.isciii.es/wp-content/uploads/2022/01/Newsletter-January-22-final_comp.pdf#25	blood pressure, hypertension, blood, biomarker, biomarkers
3 (no: Diet terms, Study methodology)	Precision Medicine Initiative All of Us	doi: 10.1038/s41467-023-38990-9 PMID: 37268629 PMCID: PMC10238525	obesity, blood pressure, blood, polygenic risk score
3 (no: Diet terms, Data methodology)	Precision Medicine Initiative All of Us	doi: 10.3390/healthcare11081138 PMID: 37107973 PMCID: PMC10137945	body weight, serum, biomarker, biomarkers
3 (no: Phenotypes terms, Diet terms)	MyCode Community Health Initiative	doi: 10.1001/jama.2018.18179 PMID: 30535219 PMCID: PMC6436530	blood, sequencing, genome, precision medicine
3 (no: Phenotypes terms, Diet terms)	Qatar Genome Programme (QGP)	doi: 10.1093/hmg/ddac243 PMID: 36168886 PMCID: PMC9990988	blood, sequencing, genome, personalized medicine
3 (no: Phenotypes terms, Diet terms)	ERA PerMed	https://erapermed.isciii.es/wp-content/uploads/2021/01/Newsletter-ERA-PerMed_final.pdf#18	image, genome, genomic, precision medicine
3 (no: Diet terms, Data methodology)	Precision Medicine Initiative All of Us	doi: 10.1016/j.ophtha.2021.10.018 PMID: 34688700 PMCID: PMC8863625	blood pressure, blood, cohort study
3 (no: Diet terms, Data types)	MyCode Community Health Initiative	doi: 10.1016/j.ajhg.2018.03.012 PMID: 29727688 PMCID: PMC5986700	hypercholesterolemia, genomic, precision medicine
3 (no: Diet terms, Data types)	Brazilian Initiative on Precision Medicine	doi: 10.1038/s41598-019-50362-2 PMID: 31554886 PMCID: PMC6761108	obesity, genome, precision medicine

3 (no: Diet terms, Data methodology)	FarGen Project	doi: 10.14814/phy2.15382 PMID: 35822425 PMCID: PMC9277514	blood pressure, blood, intervention
3 (no: Data types, Data methodology)	Health-RI	https://www.health-holland.com/project/2023/2017/periodic-use-fasting-mimicking-diet-type-2-diabetes	glycemic control, diet, intervention

Summary

Traditional literature review followed by text mining and natural language processing assisted literature review was applied to evaluate 34 precision medicine initiatives for relevance to CoDiet. A stratified list of 281 studies was produced with links and matched terms to allow the identification of relevant data for feature representation learning by WP3 and WP4 to assist with training causal machine learning models to be fine-tuned with data from WP2 when this becomes available. Synthetic data on 10,000 individuals from CIC bioGUNE has already been shared with WP3 and WP4. Most precision initiatives focus on genomic data with most studies excluded focussing on cancers. Some of the top ranked studies with most synergy to CoDiet have multi-modal and multi-omic data as will be produced by CoDiet WP2 and WP5.