Provincie Antwerpen

vito

Vlaanderen
is zorgzaam en
gezond samenleven

Feasibility study: coupling of PFAS results from human biomonitoring to health effects registered by general practitioners

PROVINCIAL INSTITUTE OF HYGIENE (PIH)
FLEMISH INSTITUTE OF TECHNOLOGICAL RESEARCH (VITO)
PARTNERS IN THE ‘PARTNERORGANISATION ENVIRONMENTAL HEALTH CARE’ OF THE FLEMISH GOVERNMENT, DEPARTMENT CARE

IN COLLABORATION WITH INTEGO

Intego

COLOFON

Proof-of-concept study on the coupling of PFAS results from human biomonitoring studies to health endpoints in the INTEGO-database.

Authors: Annelies De Decker, Ilona Gabaret, Elly Den Hond (PIH); Katleen De Brouwere, Hamid Hassen (VITO); Pieter Jansen, Endale Alemayehu Ali, Gijs Van Pottelbergh (KU Leuven).

Reviewed by Department of Care: Hanne Vandenbossche, Hilde Van De Maele

This report is a realisation of the Provincial Institute of Hygiene (PIH), the Flemish Institute of Technological Research (VITO), which are partners in the Partner Organisation Environmental Health Care, in collaboration with Intego. The Department of Care of the Flemish Government funded the study.

www.provincieantwerpen.be/aanbod/dlm/pih/onderzoek.html; www.vito.be; www.intego.be; www.vlaanderen.be/departement-zorg

ACRONYMS

ALT	Alanine Transaminase
AST	Aspartate Aminotransferase
B	Branched forms (of PFAS)
β	Intercept (in regression analysis)
BKMR	Bayesian Kernel Machine Regression (BKMR)
CPCSSN	Canadian Primary Care Sentinel Surveillance Network
COPD	Chronic Obstructive Pulmonary Disease
DAG	Directed Acyclic Graph
EMR	Electronical Medical Record (Elektronisch Medisch Dossier)
FHIR	Fast Healthcare Interoperability Resources
GGT	Gamma-Glutamyl Transferase
GP	General practitioner
HBM	Human Biomonitoring
HDL	High Density Lipoprotein
INSZ	Identificatienummer van de Sociale Zekerheid
L	Linear forms (of PFAS)
LDL	Low Density Lipoprotein
LOD	Limit of Detection
LOINC	Logical Observation, Identifiers, Names and Codes
LOQ	Limit of Quantification
NN	National Number
P	Probability
PFAS	Poly- and perfluoroalkyl substances
PFBA	Perfluorobutanoic acid
PFBS	Perfluorobutane sulfonic acid
PFDA	Perfluorodecanoic acid
PFDoA	Perfluorododecanoic acid
PFHpA	Perfluoroheptanoic acid
PFHpS	Perfluorheptane sulfonic acid
PFHxA	Perfluorohexanoic acid
PFHxS	Perfluorohexane sulfonic acid
PFNA	Perfluorononanoic acid
PFOS	Perfluorooctane sulfonic acid
PFOA	Perfluorooctanoic acid
PFPeA	Perfluoropentoic acid
PFUnA	Perfluoroundecanoic acid
PIH	Provinciaal Instituut voor Hygiëne
PO MGZ	Partner organization Environmental Health Care
RIZIV	Rijksinstituut voor Ziekte- en Invaliditeitsverzekering
SE	Standard Error
TSH	Thyroid Stimulating Hormone
U/L	Units per liter
VITO	Vlaams Instituut voor Technologisch Onderzoek
WQS	Weighted Quantile Sum Regression

X	Explanatory variable (in regression analysis)
Y	Response variable (in regression analysis)

SUMMARY

Feasibility study: coupling of PFAS results from human biomonitoring to health effects registered by general practitioners

CONTEXT

Intego is a Flemish general practice-based morbidity registration network. Participating GPs automatically have medical data extracted from their electronic medical record (EMR), which can be used for study purposes. All information is extracted through a trusted third party and is done in such a way to guarantee anonymity of medical data. The extracted data includes measurements (e.g. blood pressure, weight, …), prescriptions, lab values, diagnoses, age, gender, reimbursement status,…

In Flanders, there is a long history to set up human biomonitoring (HBM) as a surveillance network to monitor chemical exposure and the accompanying health impact in the population, both in the general population and in specific population groups living around contaminated sites (hotspot areas).

The overall aim of the current study was to investigate if it is feasible to couple results of biomarkers of exposure from HBM studies to health effects registered in the EMR of the Intego network, and to use the coupled database for research purposes to study associations between exposure to chemicals and health effects.

In the current study, the PFAS case in the neighborhood of the 3M site in Zwijndrecht was used as a practical example to study the feasibility. Data from the PFAS monitoring study of 2021 were used as test case. This PFAS monitoring study included data of 16 PFAS compounds in serum in 796 participants residing within a circle of 3km around 3M.

RESULTS

In order to evaluate the feasibility, objectives at different levels were defined. The results show that the consecutive objectives were successfully fulfilled.

1. Technical objectives

In order to allow a coupling of the data, the PFAS serum data of the participants of the PFAS monitoring study 2021 were delivered in a high quality way to the EMR of the GP. A first step was to comply with the necessary conditions to prepare a structured database with the PFAS results and the necessary parameters to be able to send lab results in a structured and secure way.

- In the initial HBM study, participants gave consent to deliver the results of their blood analysis to their GP. The RIZIV number of the GP is a necessary key to send results of a patient to the correct GP, and was therefore added to the database.
- Further, participants gave permission to use their National Insurance number (*rijksregisternummer*). A processing agreement was signed between the department of care of the Flemish government (client) and PIH (contractor, responsible for the personal data in the HBM studies) that explicitly defined that the National Insurance number could be used.

- Lab data should have a clear and uniform code to allow the correct delivery to the EMD. Therefore, new medidoc codes were assigned to each PFAS compound.

A second step was to design a protocol that allowed to send the PFAS results from PIH to the individual GPs. A macro was developed to transform the database into a structured format (.lab and .adr files) and deliver the individual HBM results in a secure way to the correct GP by using a 'Unified Messaging Module' of software deliver Corilus. This protocol was successfully tested for the three major software packages that are used in Belgium, i.e. CareConnect, Daktari and Health One. In the future, PIH will be able to send lab results to the correct recipient and to perform this in bulk.

As a final step, the Intego consortium was able to extract PFAS serum values from the EMR of the participating GPs and use these data for further statistical analysis.

2. Content objectives

In order to study exposure-effect associations, a literature review was done to select specific health endpoints that are relevant to study in relation to serum PFAS values. In a second step, it was evaluated whether it is feasible to extract these parameters from the Intego database, and these can be considered for further statistical analysis (Table I*)*. For these health endpoints, relevant confounders were defined. For the health endpoints that are difficult to extract from the Intego database (e.g. cancer), alternative databases are suggested.

Table I: List of health endpoints that were selected for further analysis

Type of health parameter	Category	Biomarker of effect or health endpoint
Diagnosis	Cardiovascular	Hypertension Ischemic events
	Kidney function	Kidney disease
	Immune system	Asthma Inflammatory bowel disease
	Hormonal system	Thyroid disease
	Pregnancy	Gestational hypertension
	Musculoskeletal system	Arthrosis
	Respiratory	Chronic bronchitis
Clinical parameter	Cardiovascular	Systolic blood pressure Diastolic blood pressure
	Respiratory	Shortness of breath
Lab value	Liver function	Alanine transaminase (ALT) Aspartate aminotransferase (AST) Gamma-Glutamyl Transferase (GGT) Bilirubin
	Kidney function	Creatinine
	Hormonal system	Thyroid Stimulating Hormone (TSH)
	Fat metabolism	Total cholesterol LDL-cholesterol HDL-cholesterol Triglycerides

3. Statistical objectives

The selected endpoints were transformed to 'case definitions'. This was done according to instructions of the Canadian Primary Care Sentinel Surveillance Network (CPCSSN), with adjustment to the Belgian context. For some variables a combination of data (e.g. prescription, lab data, diagnosis or symptom) is needed; for others (e.g. a clinical measurement) the data itself is the outcome. Outcome variables can be acute or chronic; variables types can be continuous, binary or ordinal.

As the major predictor variables, 16 PFAS variables were considered. Serum values of the following 13 PFAS compounds were used: PFBA, PFPeA, PFHxA, PFHpA, PFOA, PFNA, PFDA, PFUnA, PFDoA, PFBS, PFHxS, PFHpS and PFOS (expressed in µg/l). For PFOS, PFOA and PFHxS, both the linear (L) form and the sum of linear and branched (L+B) forms are considered. For values below the limit of quantification (LOQ) imputations were performed according to the guidelines of HBM4EU.

Covariates were also selected as predictor variables. Age, sex, smoking status, reimbursement status as proxy for socioeconomic status and body mass index (BMI) were considered as primary covariates for all exposure-response relationships. Secondary covariates were defined per health endpoint, and can optionally be added per specific model.

A statistical analysis plan was developed and different statistical scripts were programmed in R. Depending on the outcome variable, different regression models can be applied: 1) multiple linear regression for continuous outcome variables (e.g. a lab value like creatinine); 2) binary logistic regression for binary outcome variables (e.g. hypertension); 3) ordinal logistic regression for ordinal outcome variables (e.g. amount of exacerbations of COPD); 4) proportional Hazard (Cox-PH) model for the survival time of patients (e.g. did the patient suffer a myocardial infarction and when did this occur). These different models all allow to study associations between 1 PFAS compound and 1 health outcome, after adjustment for relevant covariates. To quantify the joint effect of several PFAS compounds and to identify the contribution of each exposure to the mixture, a series of multi-pollutant analysis methods was programmed, i.e. weight quantile sum regression (WQSR), G-computation quantile, Bayesian Kernel Machine Regression.

A proof-of-concept was performed on the basis of coupled data of 291 patients. Based on descriptive data, the PFAS compounds with sufficiently high detection frequencies were selected. As an example, regression models were built between PFOS, PFOA, PFHxS, PFNA and a number of continuous outcomes (total cholesterol, LDL, HDL, ALT, AST, gamma-GT, eGFR, TSH) and binary outcomes (hypertension, arthrosis), after adjustment for pre-selected covariates. Further, three mixture models were built to study associations between the 4 PFAS compounds and ALT. The results show that the models are valid since known associations (e.g. sex differences in clinical outcome) are confirmed. No significant results were observed between PFAS exposure and health endpoint. Most likely, this can be attributed to a lack of power in the current proof-of-concept.

RECOMMENDATIONS

In conclusion, this study shows that it is technically feasible to couple serum PFAS values from HBM studies and clinical data of the EMR of GPs. Based on a literature review, relevant health endpoints for PFAS exposure were selected, and translated to 'case definitions' with data from the EMR. A statistical analysis plan (and scripts in R) was developed to study exposure-response associations. The proof-of-concept showed no significant results, most likely due to a lack of power. This study holds promising perspectives for future analyses on a larger database.

From this pilot study, a number of recommendations for future research were formulated.

- **Coupling of data**: a high quality database is the basis for a good study. The following recommendations can be given:
    - In the design of HBM studies, it is important to foresee the necessary permissions in the informed consent form, i.e. the permission to transfer HBM results to the GP and the permission to use the National Insurance number of the participant.
    - Since one biomarker can be measured by different labs, and these results will be pooled in the statistical analysis, it is important to take actions to guarantee the quality of the measurements. Therefore, in the planning of HBM studies, it is recommended to select accredited laboratories only.
    - Additionally to the quality of the measurement itself, it is also important that the terminology and nomenclature is standardized. Currently, a new standard format is developed, namely 'FHIR (Fast Healthcare Interoperability Resources)'. From a research point of view, we recommend to promote this new standard.
- **Statistical analysis**: the use of routinely collected medical data for scientific research is an interesting option for reasons of efficiency and time saving. Yet, since the objectives for data collection are different, some pitfalls might occur during the process of analysis and interpretation. Therefore, it is important to include the appropriate experts and control steps throughout the project.

  The following recommendations can be given:
    - Include internal validation of covariates (e.g. effect of age, effect of gender) to assess the quality of the data.
    - Include epidemiology experts to understand the underlying data (e.g. over- or underestimation, selection bias, …).
    - Include content-related experts, e.g. toxicologists, medical specialists, etc.
    - Foresee sensitivity analyses, e.g. sub-analysis in subgroups of patients with more complete data or more follow-up data.
    - In cross-sectional analyses, consider reversed associations.
    - Take into account lag times between exposure and effect.
    - If possible, perform repeated analyses to consider different time windows in order to study PFAS compounds with a different half-life. For persistent PFAS compounds, repeated analyses can also be considered as sensitivity analyses.

SAMENVATTING

Haalbaarheidsstudie: koppeling van PFAS resultaten uit humane biomonitoring met gezondheidsdata geregistreerd door huisartsen.

KADERING

Intego is een registratienetwerk van Vlaamse huisartsen met gegevens over eerstelijnszorgmorbiditeit. Via een 'trusted third party' worden gegevens uit het elektronisch medisch dossier (EMD) geëxtraheerd en geanonimiseerd, waarna de ziektegegevens beschikbaar worden gemaakt voor wetenschappelijk onderzoek. De geëxtraheerde gegevens omvatten o.m. diagnoses, symptomen, technische metingen (bv. bloeddruk, gewicht, …), medicatie, laboratoriumwaarden, leeftijd, geslacht, terugbetalingsstatus.

In Vlaanderen bestaat er een lange traditie van het uitvoeren van humane biomonitoring (HBM) als monitoringssysteem waarbij de blootstelling aan chemische stoffen en de daarmee gepaard gaande gevolgen voor de gezondheid van de bevolking worden opgevolgd. HBM surveillance wordt zowel bij de algemene bevolking als bij specifieke bevolkingsgroepen rond vervuilde sites (hotspots) uitgevoerd.

Het algemene doel van de huidige studie was om te onderzoeken of het haalbaar is om resultaten van biomerkers van blootstelling uit HBM-studies te koppelen aan gezondheidseffecten geregistreerd in het EMD van het Intego-netwerk, en vervolgens de gekoppelde database te gebruiken om associaties tussen blootstelling aan chemische stoffen en gezondheidseffecten te bestuderen.

In deze pilootstudie wordt de PFAS-casus in de buurt van de 3M-site in Zwijndrecht als praktijkvoorbeeld gebruikt om de haalbaarheid te onderzoeken. Gegevens uit het PFAS monitoringonderzoek 2021 worden als testcase gebruikt. Dit PFAS onderzoek omvat metingen van 16 PFAS-componenten in serum van 796 deelnemers die binnen een cirkel van 3 km rond 3M woonden.

RESULTATEN

Om de haalbaarheid te beoordelen, werden doelstellingen op verschillende niveaus gedefinieerd. Uit de resultaten blijkt dat de opeenvolgende doelstellingen met succes zijn behaald.

1. Technische doelstellingen

Het doel van deze studie is een gekoppelde databank van hoge kwaliteit op te maken om dosis-effect relaties te bestuderen tussen PFAS en gezondheid. Daarom werden de PFAS-serumwaarden van de deelnemers aan het PFAS monitoringonderzoek 2021 op een veilige manier elektronisch doorgestuurd naar het EMD van de huisarts. Om de privacy te garanderen en om de verzending technisch mogelijk te maken, waren een aantal voorwaarden nodig.

- In het initiële HBM-onderzoek gaven deelnemers toestemming om de resultaten van hun bloedanalyse aan hun huisarts te bezorgen. Het RIZIV-nummer van de huisarts is een noodzakelijke sleutel om de uitslag van een patiënt naar de juiste huisarts te sturen en werd daarom aan de databank toegevoegd.
- Verder gaven de deelnemers toestemming om hun rijksregisternummer te gebruiken als unieke sleutel voor de correcte verzending naar het EMD. Er werd een verwerkersovereenkomst opgemaakt tussen het Departement Zorg van de Vlaamse overheid (opdrachtgever) en het PIH

(opdrachtnemer, verantwoordelijk voor de persoonsgegevens in de HBM-studies) waarin expliciet werd vastgelegd dat het PIH mag gebruik maken van het rijksregisternummer voor het uitvoeren van de huidige opdracht.

- Laboratoriumgegevens moeten een duidelijke en uniforme code hebben om correcte aanlevering aan en extractie uit het EMD mogelijk te maken. Daarom werd aan elke PFAS-component een nieuwe medidoc-code toegewezen.

Een tweede stap was het ontwerpen van een technisch protocol dat toeliet om de PFAS-resultaten van het PIH naar de individuele huisartsen te verzenden. Er werd een macro ontwikkeld om de database om te zetten in een gestructureerd formaat (.lab- en .adr-bestanden) en de individuele HBM-resultaten op een veilige manier aan de juiste huisarts te bezorgen door gebruik te maken van een 'Unified Messaging Module' van software leverancier Corilus. Dit protocol werd succesvol getest voor de drie grote softwarepakketten die in België door huisartsen gebruikt worden, namelijk CareConnect, Daktari en Health One. In de toekomst zal het PIH de laboratoriumresultaten van HBM-studies naar de juiste ontvanger kunnen sturen en dit in bulk kunnen uitvoeren.

Als laatste stap kon het Intego-consortium de PFAS-serumwaarden uit het EMD van de deelnemende huisartsen halen en deze gegevens gebruiken voor verdere statistische analyse.

2. Inhoudelijke doelstellingen

Om de associaties tussen blootstelling en effect te onderzoeken, werd een literatuuronderzoek uitgevoerd om specifieke gezondheidseindpunten te selecteren die relevant zijn voor PFAS (Tabel I).

Tabel I: Lijst van gezondheidseindpunten die geselecteerd werden voor verdere analyse

Type gezondheidsvariabele	Categorie	Biomerker van effect of ziekte
Diagnose	Cardiovasculair	Hypertensie Ischemisch event
	Nierfunctie	Nierziekte
	Immuunsysteem	Astma Inflammatoire darmziekte
	Hormonaal systeem	Schildklieraandoening
	Zwangerschap	Zwangerschapshypertensie
	Musculoskeletaal systeem	Artrose
	Respiratoir systeem	Chronische bronchitis
Klinische parameter	Cardiovasculair	Systolische bloeddruk Diastolische bloeddruk
	Respiratoir	Kortademigheid
Laboratoriumwaarde	Leverfunctie	Alanine transaminase (ALT) Aspartaat aminotransferase (AST) Gamma-Glutamyl Transferase (GGT) Bilirubine
	Nierfunctie	Creatinine
	Hormonaal systeem	Thyroid Stimulating Hormoon (TSH)
	Vetmetabolisme	Totale cholesterol LDL-cholesterol HDL-cholesterol Triglyceriden

In een tweede stap werd geëvalueerd of het haalbaar is om deze parameters uit de Intego-database te selecteren voor verdere statistische analyse. Voor deze gezondheidseindpunten werden tevens ook de relevante covariaten gedefinieerd. Voor de gezondheidseindpunten die moeilijk uit de Intego-database te halen zijn (bijvoorbeeld kanker), werden alternatieve databases voorgesteld.

3. Statistische doelstellingen

De geselecteerde eindpunten werden omgezet naar 'case definities'. Dit gebeurde volgens de instructies van het *Canadian Primary Care Sentinel Surveillance Netwerk* (CPCSSN), aangepast aan de Belgische context. Voor sommige variabelen is een combinatie van gegevens nodig (bijv. voorschrift, laboratoriumgegevens, diagnose of symptoom); voor anderen zijn de gegevens zelf de uitkomst (bijv. klinische meting). Uitkomsten kunnen acuut of chronisch zijn; types van variabelen kunnen continu, binair of ordinaal zijn.

In het statistisch model werden als voorspellende variabelen de 16 PFAS-componenten opgenomen, namelijk serumwaarden van 13 congeneren: PFBA, PFPeA, PFHxA, PFHpA, PFOA, PFNA, PFDA, PFUnA, PFDoA, PFBS, PFHxS, PFHpS en PFOS (uitgedrukt in µg/l); voor PFOS, PFOA en PFHxS wordt zowel de lineaire (L) vorm als de som van lineair + vertakt (L+B) in beschouwing genomen. Voor waarden onder de kwantificatielimiet (LOQ) werd een imputatie uitgevoerd volgens de richtlijnen van HBM4EU.

Covariaten werden ook geselecteerd als voorspellende variabelen. Leeftijd, geslacht, rookstatus, terugbetalingsstatus als proxy voor de socio-economische status en body mass index (BMI) werden beschouwd als primaire covariaten voor alle blootstelling-responsrelaties. Secundaire covariaten zijn per gezondheidseindpunt gedefinieerd en kunnen optioneel per specifiek model worden toegevoegd.

Er werd een statistisch analyseplan ontwikkeld en de verschillende scripts hiervoor werden geprogrammeerd in R. Afhankelijk van de uitkomstvariabele kunnen verschillende regressiemodellen worden toegepast: 1) meervoudige lineaire regressie voor continue uitkomstvariabelen (bijv. een laboratoriumwaarde zoals creatinine); 2) binaire logistische regressie voor binaire uitkomstvariabelen (bijv. hypertensie); 3) ordinale logistische regressie voor ordinale uitkomstvariabelen (bijv. aantal exacerbaties van COPD); 4) proportioneel Hazard (Cox-PH) model voor de overlevingstijd van patiënten (bijv. heeft de patiënt een hartinfarct gehad en wanneer vond dit plaats). Deze verschillende modellen maken het allemaal mogelijk om associaties tussen 1 PFAS component en 1 gezondheids-eindpunt te bestuderen, na correctie voor relevante covariaten. Om het gezamenlijke effect van verschillende PFAS-verbindingen te kwantificeren en de bijdrage van elke afzonderlijke PFAS component aan het mengsel te identificeren, werd ook een statistisch analyseplan voor mengsels (mixed exposure) opgemaakt. Hier werden drie types van analyse voorgesteld, nl. *weight quantile sum regression* (WQSR), *G-computation quantile*, *Bayesian Kernel Machine Regression*.

Op basis van gekoppelde data van 291 patiënten werd een proof-of-concept uitgevoerd. Op basis van beschrijvende gegevens werden de PFAS-verbindingen met voldoende hoge detectiefrequenties geselecteerd. Er zijn bijvoorbeeld regressiemodellen gebouwd tussen PFOS, PFOA, PFHxS, PFNA en een aantal continue uitkomstvariabelen (totaal cholesterol, LDL, HDL, ALT, AST, gamma-GT, eGFR, TSH) en binaire uitkomstvariabelen (hypertensie, artrose), met correctie voor vooraf geselecteerde covariaten. Verder werden drie mengselmodellen gebouwd om associaties tussen de 4 PFAS-verbindingen en ALT te bestuderen. De resultaten laten zien dat de modellen valide zijn, aangezien bekende associaties (bijv. verschil in klinische uitkomst volgens geslacht) worden bevestigd. Er werden geen significante resultaten waargenomen tussen de blootstelling aan PFAS en het gezondheidseindpunt. Hoogstwaarschijnlijk kan dit worden toegeschreven aan onvoldoende power omwille van lage aantallen in het huidige proof-of-concept.

AANBEVELINGEN

In deze pilootstudie werd aangetoond dat het technisch haalbaar is om serum PFAS-waarden uit HBM-studies te koppelen aan klinische gegevens uit het EMD van huisartsen. Op basis van literatuuronderzoek zijn relevante gezondheidseindpunten voor PFAS-blootstelling geselecteerd en vertaald naar 'case definities' die kunnen aangemaakt worden in het EMD. Er werd een statistisch analyseplan (en scripts in R) ontwikkeld om de associaties tussen blootstelling en respons te bestuderen. Het proof-of-concept leverde geen significante resultaten op, hoogstwaarschijnlijk als gevolg van een lage power omwille van te kleine aantallen. Deze studie biedt veelbelovende perspectieven voor toekomstige analyses op een grotere database.

Vanuit deze pilootstudie zijn een aantal aanbevelingen voor toekomstig onderzoek geformuleerd.

- **Koppelen van databanken**: een database van hoge kwaliteit is de basis voor een goed onderzoek. De volgende aanbevelingen kunnen worden gegeven:
    - Bij de design van HBM-studies is het belangrijk om in het geïnformeerde toestemmingsformulier de nodigde toestemmingen te voorzien, nl. de toestemming van de deelnemer om HBM-resultaten aan de huisarts door te geven en de toestemming om hiervoor het rijksregisternummer van de deelnemer te gebruiken.
    - Omdat eenzelfde biomerker door verschillende laboratoria kan worden gemeten en deze resultaten zullen worden samengevoegd in de statistische analyse, is het belangrijk dat de kwaliteit van de metingen gegarandeerd is. Daarom wordt aanbevolen om bij de planning van HBM-studies alleen geaccrediteerde laboratoria te selecteren.
    - Naast de kwaliteit van de meting zelf is het ook van belang dat de terminologie en nomenclatuur gestandaardiseerd zijn. Momenteel wordt een nieuw standaardformaat ontwikkeld, namelijk 'FHIR (Fast Healthcare Interoperability Resources)'. Wij raden aan om deze nieuwe standaard te promoten.
- **Statistische analyse**: het gebruik van routinematig verzamelde medische gegevens voor wetenschappelijk onderzoek is een interessante optie vanwege efficiëntie en tijdswinst. Maar aangezien de doelstellingen voor het verzamelen van gegevens verschillend zijn, kunnen zich valkuilen voordoen tijdens het analyse- en interpretatieproces. Daarom is het belangrijk om gedurende het hele project de juiste experts te betrekken en controlestappen uit te voeren. De volgende aanbevelingen kunnen worden gegeven:
    - Voer interne validatie van covariaten uit (bijv. effect van leeftijd, effect van geslacht) om de kwaliteit van de gegevens en modellen te beoordelen.
    - Betrek epidemiologische experten om de onderliggende data te begrijpen (over- of onderschatting van incidentie of prevalentie; selectie bias; ...).
    - Betrek inhoudelijke experten, bijv. toxicologen, medisch specialisten, enz.
    - Voorzie sensitiviteitsanalyses, bijv. sub-analyse in subgroepen van patiënten met meer complete gegevens of langere opvolgtijd.
    - Houd bij cross-sectionele analyses rekening met de optie van omgekeerde associaties.
    - Houd rekening met lag-tijden tussen blootstelling en effect.
    - Voer indien mogelijk herhaalde analyses uit om verschillende tijdvensters te bestuderen. Dit is enerzijds nuttig om PFAS componenten met een verschillend half-leven te bestuderen, anderzijds kan het voor persistente PFAS componenten ook als sensitiviteitsanalyse gelden.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION AND STUDY AIM 17

1.1 INTRODUCTION 17
1.2 PFAS MONITORING STUDY (2021) 17
1.3 INTEGO 18
1.4 STUDY AIM AND OBJECTIVES 18

CHAPTER 2: DATA COUPLING 20

2.1 CHOICE OF DATA COUPLING PROCEDURE 20
2.2 DATA COUPLING CRITERIA 21
2.3 DATA COUPLING 22

CHAPTER 3: RESEARCH QUESTIONS 24

CHAPTER 4: STATISTICAL ANALYSIS PLAN 26

4.1 OUTCOME VARIABLES 26
4.2 PREDICTOR VARIABLES 27
4.2.1 SELECTION OF PFAS COMPOUNDS 27
4.2.2 HANDLING OF SERUM PFAS LEVELS EXTRACTED FROM EMR TO COUPLED DATABASE 27
4.3 COVARIATES AND CONFOUNDERS 27
4.4 STATISTICAL ANALYSES PLAN AND POWER ANALYSES 28
4.4.1 SINGLE POLLUTANT ANALYSIS 28
4.4.2 MIXTURE ANALYSIS 30

CHAPTER 5: PROOF-OF-CONCEPT ANALYSES 31

5.1 DESCRIPTIVE STATISTICS 31
5.2 PRELIMINARY ANALYSES 35
5.2.1 SINGLE POLLUTANT ANALYSIS 35
5.2.1 MIXTURE ANALYSIS 43

CHAPTER 6: DISCUSSION 46

6.1 GENERAL DISCUSSION 46
6.2 STRENGTHS 46
6.3 LIMITATIONS & SOLUTIONS 47

CHAPTER 7: RECOMMENDATIONS 49

7.1 COLLECTION AND TRANSFER OF DATA 49
7.2 STATISTICAL ANALYSIS 49
7.3 BEYOND INTEGO 50

LIST OF FIGURES

Figure 1: Environment and health chain PFAS 24
Figure 2: Diagnosis status counts per outcome 35
Figure 3: Bar plots for weights of PFAS (top-left: ALT, top-right: hypertension) and scatter plot for the association between weights and outcomes (bottom-left: ALT, bottom-right: hypertension) 43
Figure 4: Weights based on g-computation (ALT: left, Hypertension: right) 44
Figure 5: Predictor-response function (Left: ALT, right: Hypertension) 45

LIST OF TABLES

Table 1: List of health parameters that were selected for further analysis 25
Table 2: Examples of grouping outcome variables 26
Table 3: Descriptive analysis for liver value related outcomes 32
Table 4: Descriptive analysis for various lab test-based health outcomes 33
Table 5: Descriptive analysis based on various diagnosis outcomes 34
Table 6: Estimates, standard error and p-values of multiple linear regression model for liver function parameters 37
Table 7: Estimates, standard error and p-values of multiple linear regression model of cholesterol-related outcomes 39
Table 8: Estimates, standard error and p-value based on multiple linear regression and logistic regression for eGFR, TSH, hypertension and arthrosis 41

CHAPTER 1: INTRODUCTION AND STUDY AIM

1.1 INTRODUCTION

Poly- and perfluoroalkyl substances (PFAS) are organic compounds with fluorine atoms bound to a carbon chain. The most well-known PFAS are perfluorooctane sulfonic acid (PFOS) and perfluoro-octanoic acid (PFOA). There is a historical and possibly also current emission of PFAS in the environment of the 3M site in Zwijndrecht (Flanders, Belgium), due to the industrial activity. In the spring of 2021, there was concern among the residents of the 3M site due to reports of increased concentrations of PFAS in the environment (soil, groundwater, eggs). This was the start of multiple studies on the exposure of PFAS in the region, both in humans through human biomonitoring (HBM) as in the environment through environmental measurements. An overview of all studies can be found at www.vlaanderen.be/pfas-vervuiling.

At the moment of writing this report, three studies in humans were conducted or started on behalf of the Flemish authorities:

- To offer a quick response to the concern of the neighboring residents, a PFAS monitoring study was set up in the summer of 2021 for citizens residing within a circle of 3km around 3M. In this study, serum PFAS levels were quantified and associations with questionnaire data (local foods, product use, and outdoor environment, life style) were studied.
- Second, a HBM study was performed in adolescents living in a circle of 5 km around 3M in the summer of 2022. Besides serum PFAS levels, biomarkers of effect and environmental measurements were performed and both source- exposure associations and dose-effect associations were studied.
- Third, a large-scale PFAS monitoring study (blood sample accompanied by questionnaire data) was offered to all local residents in a circle of 5 km around 3M. The first blood sample was taken in the spring of 2023 and is foreseen to run up to spring 2024.

The current feasibility study was set up to design a system that will allow to investigate health effects of PFAS exposure. The overall aim was to investigate if it is feasible to couple the results of the PFAS blood values in local residents to health effects registered in the Electronic Medical Records (EMR) of general practitioners (GP) connected to the Intego-network. Further, based on the scientific literature, relevant research questions were defined. Statistical methods were developed to answer these research questions. Finally, in a proof-of-concept approach, statistical analysis was performed on the coupled database. Since the database is still limited in this stage, the main aim of the interim analysis was not to generate strong conclusions, but it’s main purpose was to test the feasibility and make recommendations for future analyses when more data are available.

1.2 PFAS MONITORING STUDY (2021)

The PFAS monitoring study was conducted in 2021 by PIH and VITO, instructed and financed by the Department of Care of the Flemish government. The research area was a circle of 3 km around the 3M-site in Zwijndrecht, and included the sub-municipalities Zwijndrecht, a part of Burcht, Linkeroever and Melsele. Residents of this area could sign up voluntarily to the study; hence, the study was based on convenience sampling. A sample of residents was selected in order to have a study monitoring that lives sufficiently spread over the research area and the selected age groups (12-20 y, 21-49y, 50-99y).

In the period July-August 2021, a blood sample was taken from 796 local residents. In the serum, 16 PFAS compounds were measured. The serum PFAS levels were compared with Flemish reference values and health-based guidance values. Further, the relation between serum PFAS levels on the one hand and environmental and lifestyle factors, assessed through a questionnaire, on the other hand was studied. More details on the recruitment strategy and the results can be found in the scientific report (VITO and PIH, 2021)[1].

In the informed consent of the study, participants could indicate the wish to send their individual PFAS levels to their GP. This was done by post in the autumn of 2021. There was no experience yet at PIH or VITO with sending the results electronically to the EMR kept by the GP.

1.3 INTEGO

Intego is a Flemish general practice-based morbidity registration network. Participating GPs automatically have medical data extracted from their EMR, which can be used for study purposes. All information is extracted through a trusted third party and is done in such a way to guarantee anonymity of medical data. The extracted data includes measurements (e.g. blood pressure, weight, …), prescriptions, lab values, diagnoses, age, gender, reimbursement status,… (Delvaux et al., 2018).

As of August 2023, 130 GP practices are part of the Intego network, including over 450 GPs and containing the abovementioned medical data of over 450.000 individual patients. Intego can be viewed as an "open cohort", with participants leaving and joining the database as they move or change GPs. Data collection started in 1994, with higher quality of the data since 2000. When a new practice joins the Intego database, data can be retroactively collected, going as far back as the year 2000.

To become a registrar of Intego, a GP practice must meet the following criteria:

- The EMR software should be CareConnect. For IT purposes and to guarantee homogeneity of the extracted data, other EMR-software like Daktari or Health One cannot be used.
- The number of diagnoses registered without using keywords cannot be lower than 5%.
- The number of new diagnoses per patient per year should be higher than one.
- These parameters must be stable for at least three years.

These criteria were put in place to assure a high quality of the extracted data. Since only data coded with keywords is available, a maximal number of diagnoses should be coded. Lab values and prescriptions are generally coded automatically in the EMR.

1.4 STUDY AIM AND OBJECTIVES

The overall aim of the current study was to investigate if it is feasible to couple the results of the blood analysis of HBM studies in local residents to health effects registered in the EMR of GP's connected to the Intego network, and to investigate if it feasible to use the coupled data for research purposes investigating associations between PFAS exposure and health effects.

At the start of the feasibility study, more detailed **technical objectives** were:

- Deliver in a high quality way the PFAS data of the 'PFAS monitoring study 2021' (3 km zone around 3M) to the EMR of the GP.
- Describe how data of past and future monitoring studies can be included in the EMR in order to allow coupling with health data in the future. Recommendations can be given both for the

[1] https://assets.vlaanderen.be/image/upload/v1652873412/Bevolkingsonderzoek_PFAS_Zwijndrecht_-_Wetenschapppelijk_rapport_-_update_240222_vzdk8c.pdf

commissioning bodies that initiate the study (and define the boundary conditions) and to the researchers that perform the study.

Further, more detailed **content objectives** were:

- List relevant research questions for dose-response relationships that can be performed on the coupled database with serum PFAS levels and health effects extracted from the EMR.
- List the data needed to answer these research questions in a qualitative way.
- Make a protocol for data coupling of the PFAS data of the ‘PFAS monitoring study 2021’ to the data from the EMR with respect to the information in the informed consent of the study.

At last, more detailed **statistical objectives** were:

- Design a statistical analysis plan based on the research questions.
- In case of sufficient data: perform a proof-of-concept analysis on a pseudonymised coupled database with PFAS data and health effects extracted from the EMR data.
- Provide recommendations on the statistical analysis plan that is needed to perform a similar analysis in the future, i.e. on conditions, strengths and limitations.

CHAPTER 2: DATA COUPLING

2.1 CHOICE OF DATA COUPLING PROCEDURE

At the start of the study, two alternatives to obtain a database with coupled data on PFAS levels and health effects extracted from EMRs were identified:

- Alternative 1: sending the PFAS data from the 'PFAS monitoring study 2021' to the EMR of the GP of the participant using eHealthBox (i.e. secured electronical mailbox which makes a secured electronical communication of confidential medical data possible between Belgian actors in health care). Intego then can extract the PFAS data together with the health data of the participant from the EMRs of the GPs that are part of the Intego-network.

- Alternative 2: coupling of the database that resulted from the 'PFAS monitoring study 2021' to the Intego-database of one particular moment with as key the social security number of the participant or the name and date of birth. The first step for this option is an authorization request to the Belgian Data Protection Authority. For the merging step, a trusted third party is probably needed. In both databases, not all cases can be merged, e.g. some study participants will not have their EMR with a GP that participates in Intego; not all EMRs from patients in the region that are covered by Intego are participants in the PFAS monitoring studies.

Alternative 1 is preferred, as it has the following advantages over alternative 2:

- When PFAS data are included in the EMR of the GP, extractions of coupled serum PFAS and health effect data can be done multiple times, in the near and far future.

- No data authorisation request is needed, which saves time. For the 'PFAS monitoring study 2021', the GPs were already notified of the PFAS levels of their patients through post, since participants of the PFAS monitoring studies have given signed consent to transfer of their data to the GP.

- No trusted third party for data merging is needed.

- No data merging procedures are needed, and as such, there is less chance to have merging mistakes.

Alternative 2 has one advantage over alternative 1:

- Data on lifestyle and individual characteristics are included in the questionnaire of the studies. With alternative 2, this type of data can be combined with the serum PFAS data and coupled together to the health data extracted from the EMR. As study participants only give consent to share their blood results with the GP, this is not an option for alternative 1.

In both alternatives, the resulting database only contains data of participants from which the GP belongs to the Intego-network. The serum PFAS data from participants from which their GP does not belong to the Intego network cannot be linked to health effects registered in EMR's. To tackle this issue, it is important that a large part of GPs in the region are connected to Intego.

2.2 DATA COUPLING CRITERIA

The criteria that should be met to send HBM data from biomonitoring studies to the EMR of the GP of the participant using a secure way of data delivery are the following:

- The organisation that does the sending has the permission to use eHealth and has the right certificates for encrypted sending of medical information.
- The ethical committee has given approval for the procedure. In the ethical committee request of the 'Monitoring screening PFAS 2021', there is written: *"If the participant has given permission for this, the personal PFAS results of the blood test will be passed on to the GP. Ideally, this is done via e-Health so that the results are included directly in the patient's medical record."*

  Note for the future: only approval was asked to send PFAS results. For the future, in case it is wished to also send clinical results of the study (e.g. blood pressure, length, weight, cholesterol level etc.), this should be specified in the ethical committee request.
- In the informed consent (which is also checked by the ethical committee), approval was asked to the participant for sending his/her individual PFAS results to the GP. Also, the name and first name of the GP were asked.
- The unique social security number (*Identificatienummer van de Sociale Zekerheid, INSZ or national number, NN*) of each participant should be known. This key is a necessary key for eHealth to deliver results of an individual participant.
- Based on the name and first name of the GP, their RIZIV-number can be looked up through https://ondpanon.riziv.fgov.be/SilverPages/nl. The RIZIV-number is a necessary key for eHealth to send the results to the correct GP.
- A processing agreement was signed between the instructing and financing authority and the partner that will send the results. In this agreement, the instructing authority gives the instruction to the partner to use the social security number of the participant to inform his/her GP about the study results, and this based on the consent of the participant.

Additional criteria that should be met for extraction of PFAS serum results from the HBM database are:

- Lab data should have a clear and uniform code assigned to it (e.g. Medidoc code, LOINC code,…). Uncoded data cannot be extracted. If no code exists yet for the investigated substance, a code should be made and used by all labs, preferably uniform on an (inter)national level.

Since PFAS serum data is considered lab data and just as privacy-sensitive as other lab values, there is no ethical problem in extraction of HBM-data that is not covered by the approval of the ethical committee already. Therefore, if a patient agrees to send their PFAS results to their GP, no additional approval is needed to extract this data from the EMR.

The data that are used in the current study are limited to the region around 3M in Zwijndrecht, since in this area, a large amount of data is available. The research consortium considered the option to include data from other regions in Flanders. Data on serum PFAS levels are also available from reference populations living mainly outside PFAS affected sites (e.g. HBM in the FLEHS studies performed by the various cycles of Flemish Center of Expertise on Environment and Health (https://www.milieu-en-gezondheid.be/). However, it is not feasible for the research team to connect these results to Intego since this would demand a large effort to identify all individual GPs all over Flanders. Also, an essential condition is that the permission of the participant is present in the Informed Consent Form; this should be checked per study. Therefore, the focus stays on internal dose-effect relationships in the group of local residents participating in the Zwijndrecht region.

2.3 DATA COUPLING

The data coupling process was designed, tested and consolidated within the context of this project. PIH, in close collaboration with the GPs of KU Leuven, tested and re-tested the delivery of the HBM data to the EMR of the GP. In principle, all GPs in Intego work with the software CareConnect. However, based on experience with GPs, other widely used software packages are Daktari and Health One. Therefore, the delivery of the individual serum PFAS results was tested for these three software packages.

The following steps were taken:

- To send results using eHealth Box, two systems exist:
    - a web application that can be accessed through a webbrowser using electronical identification applications (eID/itsme).
    - a webservice that is accessible through a medical software package.

  The web application was not suitable to automate the sending of large quantities of file. Using this application was prone to make a lot of mistakes (e.g. wrong recipient, patient does not belong to the GP to which the results are sent,…). To tackle this, PIH decided to buy a license for the 'Unified Messaging Module' of one of the EMR software packages. This system has the advantage that it is suited to send a large list of individual results in bulk.

- As Intego should be able to extract the data from the EMRs working with CareConnect (delivered by Corilus), the different PFAS compounds had to be uniformly and uniquely coded. Therefore, 'medidoc codes' were generated by software deliverer Corilus on our request, and they made them available on their website for the other EMR software packages (https://labcodes.corilus.be/).

- In order to send the data via eHealthBox, an essential step was to automate the formatting of the result file according to the coding protocol delivered by the software deliverer, namely the 'medidoc coding'. Following the software deliverer, every other software package should be able to receive this protocol in a readable way.

- As a first step, a database was created with the individual PFAS results and identification of each participant coupled to GP and the GPs coordinates. The structured database was designed according the criteria of the coding protocol.

- Starting from this structured database, a macro was developed to transform the structured database into .lab-files (all details of the sending, results inclusive, that are structured following the coding protocol) and .adr-files (contains only a code for the recipient). These files are delivered to the Unified Messaging Module, which then send all files to the correct recipient.

- Testing with the most frequently used software packages if the .lab-files and .adr-files can be read in and visualized correctly.
    - Software package 'CareConnect' could read in the medidoc coding perfectly (e.g. date of blood extraction, visualisation as a lab result, colouring results out of the normal range red,…).
    - Software package 'Daktari' was not able to read the files as a lab result, and hence, reference values and identifying abnormal results was not possible. Further, Daktari did not transform the medidoc codes for the PFAS components into the actual names of the PFAS components. This issue was tackled by adding comments to each PFAS component, and by reporting the name of the PFAS component in the comment.
    - Software package 'Health One' experienced a similar problem: the system could receive text files, but was not able to read the files as a lab result. Hence, reference values were lacking which made interpretation difficult. Furthermore, the date of sampling was missing. This issue was tackled in the same way as for Daktari. Our remarks to improve the system in the future were sent to the software developer, and will be followed up in future projects.
- After the test phase, all results were sent in medidoc coding format using the Unified Messaging module. Although the results were not perfectly visualised in software package 'Daktari' and 'Health One', GPs were satisfied with the current format.
- Based on information of a contact person from a clinical lab, it is a known challenge in the field that every software package uses its own coding rules. Also, clinical labs often choose their own coding for new compounds (such as PFAS), which makes it difficult to extract these lab results based on EMRs. A new standard format is therefore developed, namely 'FHIR' (Fast Healthcare Interoperability Resources). As this is a very recent evolution, and not all software packages for EMRs have implemented the FHIR format, the new standard format could not be used within the current project. However, it is important to follow up the progress of this new system, and when this standard format is available, explore whether FHIR coding can be implemented in the protocol to send lab results to the EMR.
- Data is extracted from the EMR of participating GPs automatically on a weekly basis. It arrives in the data warehouse through HealthData and can be made available in the Intego research environment on request, where analyses are possible using R software.

In conclusion, exploring how to code HBM results in order to send individual results of participants to the different EMR software packages of GPs, was a very time consuming process. There is not one universal coding system that allows an optimal delivery of the data by the three main software packages. Also, there is a lack of transparent information on the procedures. Yet, PIH was able to develop a protocol in different steps: 1) acquire a license for the 'Unified Messaging Module' to send results in a secure manner to the EMRs of the GPs; 2) develop a structured database; 3) have medidoc codes created for PFAS in serum and 4) develop a protocol to send lab results to the correct recipient and perform this in bulk. The transfer of PFAS serum values to the EMR of the GP was done highly successful for CareConnect, and sufficiently successful for Daktari and Health One. The Intego consortium was able to extract the PFAS serum values from the EMR and use these data for further statistical analysis (see chapter 4).

CHAPTER 3: RESEARCH QUESTIONS

The overall goal of the study is to investigate the feasibility to study exposure-response associations via a coupled database of Intego. More specifically, the research question is whether PFAS serum levels generated in 'PFAS monitoring study 2021' can be linked to health effects extracted from the EMR in the region around 3M in Zwijndrecht. In the current study, a pilot is initiated. When successful with regard to methodology, this exercise can be repeated in the future on larger datasets. Since a large-scale study is ongoing, it can be expected that more data will be available within a reasonable time frame.

Figure 1 shows the 'environment and health chain' starting from environmental exposure to pollutants and ending with health effects. The coupled Intego database will not have information on the environmental exposure, but will have information on levels of serum PFAS levels of local residents (i.e. biomarker of exposure), on intermediate effect markers (biomarkers of effect) and on health endpoints.

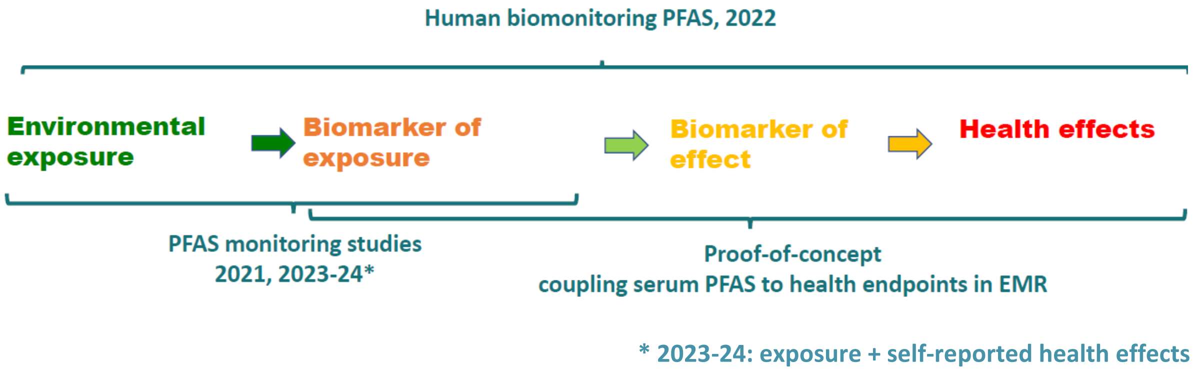
Human biomonitoring PFAS, 2022
Environmental exposure
Biomarker of exposure
Biomarker of effect
Health effects
PFAS monitoring studies 2021, 2023-24*
Proof-of-concept coupling serum PFAS to health endpoints in EMR
* 2023-24: exposure + self-reported health effects

Figure 1: Environment and health chain PFAS

In order to study **exposure-effect relationships**, a literature review was done to define specific endpoints that are of interest to study in relation to PFAS, so that relevant **research questions** can be defined.

A literature review was performed to select relevant health endpoints that can be studied in relations to biomarkers of exposure to PFAS. Firstly, a review of the Centre for Disease Control and Prevention of the U.S. was used as a basis (ATSDR, 2021). Secondly, individual scientific papers on exposure-effect associations in known PFAS hotspots areas, i.e. the Veneto region (Canova et al., 2020; Catelan et al., 2021; Gallo et al., 2022) and Ronneby region (Hammarstrand et al., 2021; Pitter et al., 2020; Xu et al., 2020a, 2020b). Different types of health parameters were considered, i.e. biomarkers of effect (lab measurements such as thyroid hormones or clinical outcomes such as blood pressure) that reflect intermediate effect parameters on the one hand, and diagnoses (diseases such as hypertension) that reflect health endpoints on the other hand.

In a first step, health parameters that may be associated with PFAS exposure were selected from literature.

In a second step, the possibility to extract this health parameter from the Intego database and use as a health outcome parameter in the statistical analysis for the current PFAS study was evaluated.

Finally, the research team – as a form of expert opinion – assigned a score to each health outcome to decide on the selection for the statistical analysis plan. The following scores were given:

- Score = 0: there is no evidence from literature for an association with PFAS and this health parameter from literature or within the Intego database, this health parameter is not suited for the current study design.
- Score = 1: there is consistent evidence from literature for an association with PFAS and it is possible to study this parameter in the Intego database.
- Score = 2: there is weak evidence for an association with PFAS and/or it is difficult to extract this parameter from the Intego database and/or other databases are more suitable for studying the health parameter (for example: cancer)

Table 1 summarizes the selected health parameters that were scores as '1'; these variables were taken up for further elaboration, i.e. a selection of confounders and covariates is done in the statistical analysis plan. If it is concluded that the data analysis of the Intego database is possible and relevant, the health parameters with a score of 2 may be considered in a further analysis. The information on the consistency of evidence, suitability of the data and scoring per health parameter is available as a Google Doc (annex 1).

Table 1: List of health parameters that were selected for further analysis

Type of health parameter	Category	Biomarker of effect or health endpoint
Diagnosis	Cardiovascular	Hypertension Ischemic events
	Kidney function	Kidney disease
	Immune system	Asthma Inflammatory bowel disease
	Hormonal system	Thyroid disease
	Pregnancy	Gestational hypertension
	Musculoskeletal system	Arthrosis
	Respiratory	Cronic bronchitis
Clinical parameter	Cardiovascular	Systolic blood pressure Diastolic blood pressure
	Respiratory	Shortness of breath
Lab value	Liver function	Alanine transaminase (ALT) Aspartate aminotransferase (AST) Gamma-Glutamyl Transferase (GGT) Bilirubin
	Kidney function	Creatinine
	Hormonal system	Thyroid Stimulating Hormone (TSH)
	Fat metabolism	Total cholesterol LDL-cholesterol HDL-cholesterol Triglycerides

CHAPTER 4: STATISTICAL ANALYSIS PLAN

4.1 OUTCOME VARIABLES

For the selected outcome variables, we need to create so-called “case definitions”. In other words, we need to define which patients in the Intego database have each outcome. These case definitions were created based on the case definitions from the Canadian Primary Care Sentinel Surveillance Network, which has validated case definitions for many outcomes (CPCSSN, 2023). These definitions were adjusted to the Belgian context.

Several types of data were used to create a case definition. Combining several types of data (e.g. prescriptions, lab data, diagnosis or symptom codes) to create a definition can sometimes increase the sensitivity and specificity. For some outcomes such as lab data, no case definition needs to be made since the lab value itself is the outcome. Intego has a list with all outcome definitions available. Additional outcome definitions can be created if necessary for future projects. For example, chronic kidney disease can be defined as: code (U99; N18.9) in medical history (Active) or in Evaluation during consultation in the past year, OR 2 measurements of eGFR <60 with a minimum of three months in between.

After a case definition was constructed for every outcome, the outcome variables were divided 1) into data types (continuous, binary or ordinal) and 2) into acute versus chronic outcomes. This is important for the development of a statistical plan. This is also included in the outcome list of Intego.

When all outcome variables were properly defined based on the data types in Intego, an overview was made to make clear which endpoints could be grouped together in case of insufficient power. For example, there might be insufficient myocardial infarctions within a given time period to make sound conclusions. As alternative, we might group myocardial infarctions together with other ischemic disease (e.g. stroke, peripheral arterial disease,…) to increase the amount of outcomes and thus increase our statistical power. An example of a grouping is given in Table 2.

Table 2: Examples of grouping outcome variables

Disease	Group 1	Group 2	Group 3
K74: Ischemic heart disease with angina	Ischemic heart disease	Ischemic cardiovascular disease	Cardiovascular disease
K75: Acute myocardial infarction			
K76: Ischemic heart disease without angina			
K89: Temporary cerebral ischemia	Ischemic cerebrovascular disease		
K90: Stroke			
K92: Atherosclerosis/Disease of periferal arteries	Periferal ischemic disease		
K77: Heart failure	Heart failure	Heart failure	

4.2 PREDICTOR VARIABLES

4.2.1 Selection of PFAS compounds

In the PFAS monitoring study 2021, 16 PFAS compounds were measured in serum samples of the participants, i.e. 13 PFAS compounds: PFBA, PFPeA, PFHxA, PFHpA, PFOA, PFNA, PFDA, PFUnA, PFDoA, PFBS, PFHxS, PFHpS and PFOS. The majority of these compounds occur mainly as linear compounds (L); in contrast, PFOS, PFOA and PFHxS occur both in linear (L) and branched (B). From an analytical point of view, concentration of linear (L) forms and concentration of linear+branched (L+B) forms are measured. For these 3 compounds, both the linear forms (PFOS (L), PFOA (L), PFHxS (L)) and the sum of linear + branched forms (PFOS (L+B), PFOA (L+B), PFHxS (L+B)) are reported. Hence, in total 16 PFAS compounds can be studied as predictor variables.

Since this is a feasibility study, it was not possible to develop and run statistical analysis for all measured 16 PFAS compounds because of capacity reasons. Therefore, a stepwise approach was applied. Firstly, a descriptive analysis is produced for all 16 PFAS compounds. Based on the detection frequencies and the proportions of the health outcome parameters, it is decided which analysis is possible. The current analysis is a proof-of-concept analysis, and therefore we want to demonstrate the feasibility of the method for different types of outcome parameters (continuous / binary). We do not have to ambition to produce valid results with respect to exposure-response associations due to the limited numbers.

This feasibility study is a test case on the PFAS monitoring study 2021. The data extraction in the current study was performed on August 3rd, 2023. In this stage, the majority of the results came from the ‘PFAS monitoring study 2021’. However, a limited number of samples may originate from other studies (e.g. large-scale monitoring study 2023-24 or results from analyses that were requested by individual GP’s). In the future, the study might be extended to PFAS compounds from other studies, to additional chemical compounds and/or other health outcome parameters.

4.2.2 Handling of serum PFAS levels extracted from EMR to coupled database

The serum PFAS values (including number of significant digits)(expressed in µg/l) available in the EMR are transferred to the coupled database. For values below the limit of quantification (LOQ), a post treatment of data is needed for statistical analysis. Hereto, values below LOQ are recalculated via random imputations if at least 30% of the values were above LOQ (in accordance with the method used in HBM4EU (Ottenbros et al., 2021). First, a truncated log-normal distribution was fitted through the observed values (namely, the values above the LOQ). This resulted in the estimation of the mean and standard deviation of the log-normal distribution of all measurements (below and above the LOQ). Values were then randomly imputed for the measurements below the LOQ, drawn between 0 and the limit from the log-normal distribution with the estimated mean and standard deviation.

4.3 COVARIATES AND CONFOUNDERS

In addition to PFAS, a multitude of other factors can influence the investigated health outcomes. Covariates, which have an influence on the outcome, and confounders, which have an influence on both the exposure (PFAS) and the outcome, should therefore be included in the statistical analysis to increase the validity of our results.

For every outcome, a list of possible covariates and confounders was made based on the medical and scientific knowledge of researchers from PIH, VITO and Intego and a non-systematic literature review.

Covariates were divided into primary covariates (covariates that are always incorporated in the statistical analyses) and secondary covariates (covariates that are added to the model to see whether their addition adds to the robustness of our analyses). Confounders were also always added to the model.

List of primary covariates; they are all considered as confounders:

- Age
- Sex
- Smoking status
- Reimbursement status, as proxy for socioeconomic status
- Body Mass Index (BMI)
- Problematic alcohol use (liver values)

Note that age, gender, smoking status, socioeconomic status and BMI are selected as primary covariates, but also are all considered as (possible) confounders, since they have a potential influence on the health outcome and are possibly correlated with serum PFAS levels according to several studies (annex 1). There is evidence for an association, but the direction of causality is not always clear.

List of secondary covariates:

- Problematic alcohol use (for hypertension, ischemic events, cholesterol)
- Diseases that increase blood pressure, e.g. obstructive sleep apnea syndrome (for hypertension)
- Stress and fear (for hypertension, dyspnea)
- Hypertension (for ischemic events, kidney disease)
- Cholesterol levels (for ischemic events, kidney disease)
- Diabetes mellitus (for ischemic events, kidney disease)
- Chronic viral hepatitis (for liver values)
- Preterm birth (for asthma)
- Lung infections at a young age (for asthma)
- Joint diseases (for arthrosis)

Some other covariates are not available in Intego, i.e.

- Physical activity
- Dietary factors
- Non-problematic alcohol use
- Family history of certain diseases (e.g., family history of cardiovascular events, asthma,…)
- Job and working environment

Finally, it is important to note that there are other environmental exposures that can have negative effects on health (e.g. heavy metals, air pollution,…). It is possible that exposure to other environmental factors is also higher in regions where PFAS pollution is present. Since other environmental exposure cannot be accounted for in this feasibility study on the Intego database, this should be considered when reporting our findings.

4.4 STATISTICAL ANALYSES PLAN AND POWER ANALYSES

4.4.1 Single pollutant analysis

The statistical test used for each outcome depends on the data type of the outcome variable. The exposure variables (PFAS-values) are in general continuous, while our selected outcomes can be binary, ordinal or continuous. Binary outcomes are further divided into acute or chronic diseases. The different analytic methods for each outcome type are listed below.

Multiple linear regression: Statistical method that is used to predict continuous outcome variables based on the value of two or more either continuous or dichotomous covariates. This method is used to analyze continuous outcomes, e.g. lab values like creatinine.

$Yi=\beta0+\beta1xi1+\beta2xi2+\ldots+\beta kxik+\epsilon i$ For i= 1, 2, . . .n number of observation

Where:

Y_i is response variable
X_{ik} are explanatory variables (covariates)
β_0 is intercept
B_k slope coefficients for explanatory variables
ϵ_i Model's error term (residual)

The main assumptions of multiple linear regression are normality, linearity and homoscedasticity (constant variance).

Binary Logistic regression: Used in a situation where the outcome variable is binary with one or more explanatory variables. This technique is used to analyze chronic binary outcomes, where those who have a diagnosis in their EMR are counted as 1 and those who don’t as 0. For example, hypertension can be analyzed this way.

$\log(p/(1-p))=\beta0+\beta1x1+\beta2x2+\ldots+\beta kxk$

Where:

P is probability that Y=1 given x
X_k are explanatory variables (covariates)
β_0 is intercept
B_k slope coefficients for explanatory variables

Ordinal Logistic regression: used to model the relationship between an ordinal response variable and one or more explanatory variables. An example of an ordinary value is the amount of exacerbations of asthma or COPD in a well-defined timeframe.

$\log(P(Y\leq j)/P(Y>j)) = \beta j0+\beta1x1+\beta2x2+\ldots+\beta kxks$

Proportional Hazard (Cox-PH) Model: Used to investigate the association between the survival time of patients and one or more predictor variables. This can be used to analyze acute binary outcomes, such as myocardial infarctions (i.e. did the patient suffer a myocardial infarction and if so, when did this occur). This model was not applied in the current study, but is added for the sake of completeness.

$h(t) = ho(t)\exp(\beta 1x1 + \beta 2x2 + \cdots + \beta kxk)$

Where:

t represents the survival time
h(t) is the hazard function be to determined based on a set of k predictor variables x1, x2, … xk.
β_k are effect sizes for explanatory variables
$h_0(t)$ is the baseline hazard which corresponds to the h=value ot hazard when all xi is equal to zero.

4.4.2 Mixture analysis

To quantify the joint effect of several correlated exposure chemicals and to identify the contribution of each exposure to the mixture, a series of multi-pollutant mixture analysis methods might be employed. PFAS compounds with at least 70% detected can considered for a mixture analysis.

To estimate the joint effect of selected PFAS compounds on the health outcomes and to identify the predominant contributor, a Weighted Quantile Sum Regression (WQS) is performed. The study sample is randomly divided into a training (40%) and validation (60%) dataset. Using the training dataset, each chemical is scored into quartiles and the total quantile score is created for individual participants. Empirical weights of each PFAS compound in the mixture are estimated through 'bootstrapping' and these weights are used to create WQSR scores representing the whole mixture. Bootstrapping is a statistical procedure that re-samples a single dataset to create many simulated samples.

To accommodate the bidirectional associations (both positive and negative), a Quantile-based g-Computation is performed, estimating the parameters of a marginal structural model, rather than a standard regression.

To examine the consistency of results and accommodate non-linear associations between PFAS compounds and the health outcome, additional mixture analysis, a Bayesian Kernel Machine Regression (BKMR) is conducted using the 'BKMR' package in R. BKMR helps to identify the uncertain exposure-outcome relationship, linear or non-linear, by non-parametric method (kernel function) and then evaluates the exposure mixtures. Furthermore, this method is also helpful to identify interaction between PFAS compounds.

All multi-pollutant models are adjusted for the same set of covariates, selected based on their influence on association of exposures with health outcomes, biological relevance, and the directed acyclic graph (DAG).

CHAPTER 5: PROOF-OF-CONCEPT ANALYSES

5.1 DESCRIPTIVE STATISTICS

There are 291 patients with at least one type of PFAS compound quantified in their EMR. This number mainly includes participants of the PFAS monitoring study 2021 (with a total number of 796) and a limited number of participants from other studies (e.g. large-scale monitoring study 2023-24) or individual initiatives (requested by the GP).

As a result of combining separate PFAS compounds with patient characteristics and health outcomes, the sample size was reduced.

Table 3, Table 4 and Table 5 summarise the descriptive results for PFAS together with liver values and various lab-based outcomes. The values in the table represent the total number of patients with a value for both the PFAS compound and the liver value. Between brackets, the percentage of measurements for the PFAS values below the Limit of Detection (LOD), below the Limit of Quantification (LOQ), and the total percentage of detection (quantified or not) is displayed. Combinations marked in orange indicate that there are insufficient measurements above LOQ (for PFAS compound of interest) to enable data imputation (see 4.2.2) and to perform analyses. For the combinations marked in yellow, we used imputation techniques to impute the PFAS values less than LOQ. PFAS compounds with 100% of detected values or with no values less than LOQ were LB-PFOS, L-PFOA, PFNA, LB-PFHxS, L-PFHxS, LB-PFOA. These can be used for analyses without the need for imputation techniques (marked green in table 5.1 and 5.2).

Figure 2 summarizes the number of patients diagnosed with specific health outcomes and those not. Out of 68 patients, only 15 of them were diagnosed with arthrosis, whereas only 1 patient was diagnosed with chronic obstructive pulmonary disease (COPD) and ischemic disease. About 20 patients were diagnosed with hypertension. Due to the fact that the number of events (diagnosed) was quite small for many outcomes, further model analyses were made for hypertension and arthrosis outcomes only, as these were the more frequently occurring diagnoses and we wanted to perform analyses with binary outcome data for the purpose of this feasibility study.

Results in Table 5 are the same descriptive summary as Table 3 and Table 4, except in this case the outcomes are diagnosis-based and are binary. For most of the outcomes across the PFAS values the total sample size is 68. Further model analysis is possible for hypertension and arthosis outcomes with PFAS (L), PFOS (L+B), PFOA (L), PFOA (L+B), PFHxS (L), PFHxS (L+B) and PFNA without imputation.

Table 3: Descriptive analysis for liver value related outcomes

PFAS compound	**Outcome**				
	Liver Values				
	ALT	**AST**	**Gamma G-T**	**Alkaline phosphates**	**Billirubin**
	Measured count (<LOD %, <LOQ %, Detected %)				
Perfluorobutanoic acid (PFBA)	62 (19, 67, 81)	61 (20, 66, 80)	62 (19, 68, 81)	59 (19, 68, 81)	27 (22, 63, 80)
Perfluoropentoic acid (PFPeA)	62 (21, 79, 79)	61 (21, 79, 79)	62 (21, 79, 79)	59 (20, 80, 80)	27 (26, 74, 74)
Perfluorohexanoic acid (PFHxA)	62 (21, 79, 79)	61 (21, 79, 79)	62 (21, 79, 79)	59 (20, 80, 80)	27 (26, 74, 74)
Perfluoroheptanoic acid (PFHpA)	62 (18, 55, 82)	61 (18, 56, 82)	62 (18, 55, 82)	59 (19, 58, 81)	27 (19, 59, 81)
Perfluorooctane sulfonic acid (PFOS)(L+B)	69 (0, 0, 100)	63 (0, 0, 100)	69 (0, 0, 100)	61 (0, 0, 100)	29 (0, 0, 100)
Perfluorooctanoic acid (PFOA)(L)	62 (0, 0, 100)	61 (0, 0, 100)	62 (0, 0, 100)	59 (0, 0, 100)	27 (0, 0, 100)
Perfluorononanoic acid (PFNA)	62 (0, 0, 100)	61 (0, 0, 100)	62 (0, 0, 100)	59 (0, 0, 100)	27 (0, 0, 100)
Perfluorodecanoic acid (PFDA)	62 (6, 5, 94)	61 (7, 5, 93)	62 (6, 5, 94)	59 (7, 5, 93)	27 (11, 7, 89)
Perfluoroundecanoic acid (PFUnA)	62 (16, 37, 84)	61 (16, 38, 84)	62 (16, 37, 84)	59 (15, 37, 85)	27 (19, 41, 81)
Perfluorododecanoic acid (PFDoA)	62 (21, 79, 79)	61 (21, 79, 79)	62 (21, 79, 79)	59 (20, 80, 80)	27 (26, 74, 74)
Perfluorobutane sulfonic acid (PFBS)	62 (21, 79, 79)	61 (21, 79, 79)	62 (21, 79, 79)	59 (20, 80, 80)	27 (26, 74, 74)
Perfluorohexane sulfonic acid (PFHxS)(L+B)	62 (0, 0, 100)	61 (0, 0, 100)	62 (0, 0, 100)	59 (0, 0, 100)	27 (0, 0, 100)
Perfluorohexane sulfonic acid (PFHxS)(L)	62 (0, 0, 100)	61 (0, 0, 100)	62 (0, 0, 100)	59 (0, 0, 100)	27 (0, 0, 100)
Perfluorheptane sulfonic acid (PFHpS)	62 (16, 10, 84)	61 (16, 10, 84)	62 (16, 10, 84)	59 (15, 10, 85)	27 (22, 7, 78)
Perfluorooctanoic acid (PFOA)(L+B)	62 (0, 0, 100)	61 (0, 0, 100)	62 (0, 0, 100)	59 (0, 0, 100)	27 (0, 0, 100)
Perfluorooctane sulfonic acid (PFOS)(L)	67 (7, 0, 93)	67 (7, 0, 93)	66 (6, 0, 94)	63 (6, 0, 94)	32 (13, 0, 87)

The colors represent our ability to do analyses. Orange means that no analysis is possible (this is the case for less frequently measured PFAS). Yellow means analysis is possible after imputation. Green means analysis is possible without the need for imputation techniques

Table 4: Descriptive analysis for various lab test-based health outcomes

PFAS compound	**Outcome**					
	Tot. cholesterol	**LDL**	**HDL**	**Triglyceride**	**TSH**	**eGFR**
	Measured count (<LOD %, <LOQ %, Detected %)					
Perfluorobutanoic acid (PFBA)	61 (16, 70, 84)	60 (15, 72, 85)	61 (15, 72, 85)	61 (15, 72, 85)	53 (15, 75, 85)	62 (19, 68, 81)
Perfluoropentoic acid (PFPeA)	61 (16, 100, 0)	60 (17, 82, 83)	61 (16, 82, 84)	61 (16, 82, 84)	53 (17, 83, 83)	62 (21, 77, 79)
Perfluorohexanoic acid (PFHxA)	61 (16, 100, 0)	60 (17, 83, 83)	61 (16, 84, 84)	61 (16, 84, 84)	53 (17, 83, 83)	62 (21, 77, 79)
Perfluoroheptanoic acid (PFHpA)	61 (16, 57, 84)	60 (15, 58, 85)	61 (15, 59, 85)	61 (15, 59, 85)	53 (15, 60, 85)	62 (18, 56, 82)
Perfluorooctane sulfonic acid (PFOS)(L+B)	65 (0, 0, 100)	62 (0, 0, 100)	65 (0, 0, 100)	63 (0, 0, 100)	53 (0, 0, 100)	67 (0, 0, 100)
Perfluorooctanoic acid (PFOA)(L)	61 (0, 0, 100)	60 (0, 0, 100)	61 (0, 0, 100)	61 (0, 0, 100)	53 (0, 0, 100)	62 (0, 0, 100)
Perfluorononanoic acid (PFNA)	61 (0, 0, 100)	60 (0, 0, 100)	61 (0, 0, 100)	61 (0, 0, 100)	53 (0, 0, 100)	62 (0, 0, 100)
Perfluorodecanoic acid (PFDA)	61 (5, 7, 95)	60 (3, 7, 93)	61 (5, 7, 95)	61 (3, 7, 97)	53 (4, 6, 96)	62 (6, 6, 94)
Perfluoroundecanoic acid (PFUnA)	61 (11, 39, 89)	60 (12, 38, 88)	61 (11, 38, 89)	61 (11, 38, 89)	53 (11, 40, 89)	62 (16, 35, 84)
Perfluorododecanoic acid (PFDoA)	61 (18, 82, 82)	60 (17, 83, 83)	61 (16, 84, 84)	61 (16, 84, 84)	53 (17, 83, 83)	62 (21, 77, 79)
Perfluorobutane sulfonic acid (PFBS)	61 (16, 84, 84)	60 (17, 83, 83)	61 (16, 84, 84)	61 (16, 84, 84)	53 (17, 83, 83)	62 (21, 77, 79)
Perfluorohexane sulfonic acid (PFHxS)(L+B)	61 (0, 0, 100)	60 (0, 0, 100)	61 (0, 0, 100)	61 (0, 0, 100)	53 (0, 0, 100)	62 (0, 0, 100)
Perfluorohexane sulfonic acid (PFHxS)(L)	61 (0, 0, 100)	60 (0, 0, 100)	61 (0, 0, 100)	61 (0, 0, 100)	53 (0, 0, 100)	62 (0, 0, 100)
Perfluorheptane sulfonic acid (PFHpS)	61 (11, 11, 89)	60 (12, 10, 88)	61 (11, 11, 89)	61 (7, 7, 93)	53 (11, 13, 89)	62 (16, 10, 84)
Perfluorooctanoic acid (PFOA)(L+B)	61 (0, 0, 100)	60 (0, 0, 100)	61 (0, 0, 100)	61 (0, 0, 100)	53 (0, 0, 100)	62 (0, 0, 100)
Perfluorooctane sulfonic acid (PFOS)(L)	65 (6, 0, 94)	61 (0, 0, 100)	64 (6, 0, 94)	65 (5, 0, 95)	58 (9, 0, 89)	67 (7 0, 93)

The colors represent our ability to do analyses. Orange means that no analysis is possible (this is the case for less frequently measured PFAS). Yellow means analysis is possible after imputation. Green means analysis is possible without the need for imputation techniques.

Table 5: Descriptive analysis based on various diagnosis outcomes

PFAS compound	Binary outcome					
	Hypertension	Arthrosis	Asthma	Ischemic events	COPD	Thyroid disease
	Total Number of patients (<LOD %, <LOQ %, Detected %)					
Perfluorobutanoic acid (PFBA)	68 (19, 69, 81)	68 (19, 69, 81)	68 (19, 69, 81)	68 (19, 69, 81)	68 (19, 69, 81)	68 (19, 69, 81)
Perfluoropentoic acid (PFPeA)	68 (21, 78, 79)	68 (21, 78, 79)	68 (21, 78, 79)	68 (21, 78, 79)	68 (21, 78, 79)	68 (21, 78, 79)
Perfluorohexanoic acid (PFHxA)	68 (21, 78, 79)	68 (21, 78, 79)	68 (21, 78, 79)	68 (21, 78, 79)	68 (21, 78, 79)	68 (21, 78, 79)
Perfluoroheptanoic acid (PFHpA)	68 (18, 54, 82)	68 (18, 54, 82)	68 (18, 54, 82)	68 (18, 54, 82)	68 (18, 54, 82)	68 (18, 54, 82)
Perfluorooctane sulfonic acid (PFOS)(L+B)	75 (0, 0, 100)	75 (0, 0, 100)	75 (0, 0, 100)	75 (0, 0, 100)	75 (0, 0, 100)	75 (0, 0, 100)
Perfluorooctanoic acid (PFOA)(L)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)
Perfluorononanoic acid (PFNA)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)
Perfluorodecanoic acid (PFDA)	68 (6, 7, 94)	68 (6, 7, 94)	68 (6, 7, 94)	68 (6, 7, 94)	68 (6, 7, 94)	68 (6, 7, 94)
Perfluoroundecanoic acid (PFUnA)	68 (16, 37, 84)	68 (16, 37, 84)	68 (16, 37, 84)	68 (16, 37, 84)	68 (16, 37, 84)	68 (16, 37, 84)
Perfluorododecanoic acid (PFDoA)	68 (21, 79, 79)	68 (21, 79, 79)	68 (21, 79, 79)	68 (21, 79, 79)	68 (21, 79, 79)	68 (21, 79, 79)
Perfluorobutane sulfonic acid (PFBS)	68 (21, 78, 79)	68 (21, 78, 79)	68 (21, 78, 79)	68 (21, 78, 79)	68 (21, 78, 79)	68 (21, 78, 79)
Perfluorohexane sulfonic acid (PFHxS)(L+B)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)
Perfluorohexane sulfonic acid (PFHxS)(L)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)
Perfluorheptane sulfonic acid (PFHpS)	68 (16, 10, 84)	68 (16, 10, 84)	68 (16, 7, 84)	68 (16, 7, 84)	68 (16, 7, 84)	68 (16, 7, 84)
Perfluorooctanoic acid (PFOA)(L+B)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)
Perfluorooctane sulfonic acid (PFOS)(L)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)	68 (0, 0, 100)

The colors represent our ability to do analyses. Orange means that no analysis is possible (this is the case for less frequently measured PFAS). Yellow means analysis is possible after imputation. Green means analysis is possible without the need for imputation techniques.

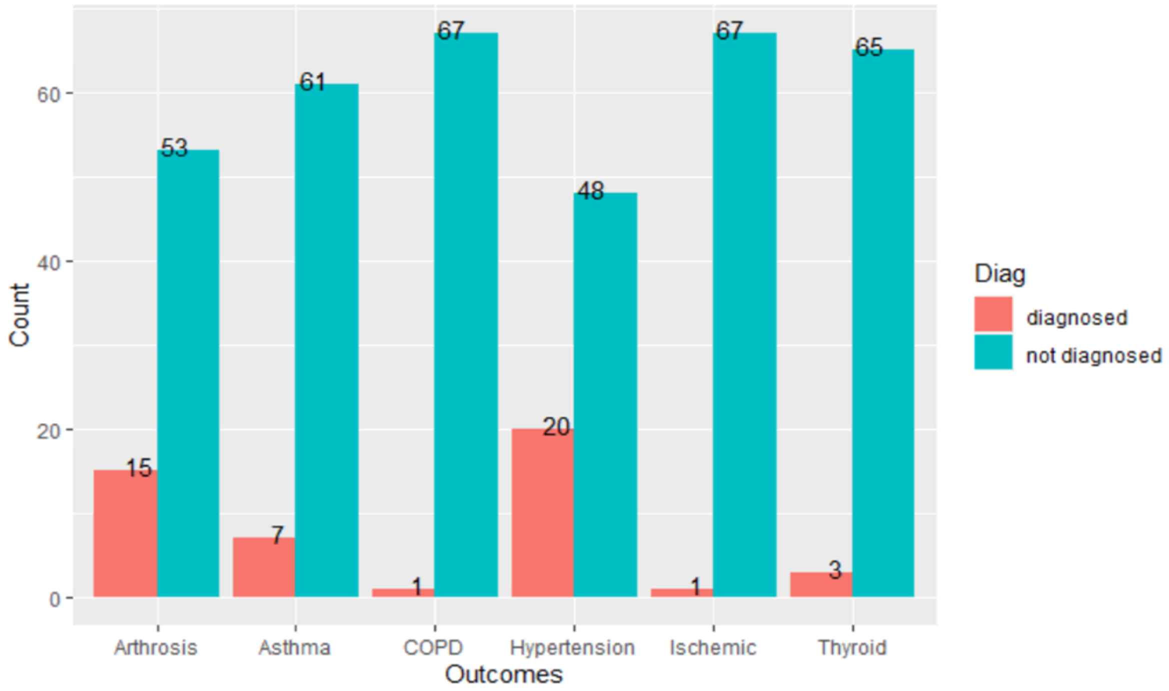

Figure 2: Diagnosis status counts per outcome

5.2 PRELIMINARY ANALYSES

5.2.1 Single pollutant analysis

Table 6 summarizes the results of our analysis of different PFAS compounds on liver values, obtained from multiple linear regression. There is no evidence for an association between any PFAS compound and ALT, AST, Gamma G-T, or Alkaline Phosphate (p>0.05).

For the model with PFNA as exposure, there is a significant association of BMI and ALT (p=0.04). As BMI increased by one kg/m^2, the amount of ALT in blood is increased by 0.03 U/L when all other covariates stay constant, which is a weak association. Compared to those were not smoking, the amount of ALT in blood of patients with unknown smoking status is lower by about 5.49 U/L. There is no statistically significant association between age, sex and smoking status, and ALT for all the other models.

Similarly, there is no association between any of the covariates and AST, except a weak association with age in the model for PFOA .

There is a statistically significant association between sex and Gamma G-T in all the models (p<0.05). For example, considering the model with PFOS (L+B) as exposure, males on average have an 11 U/L higher value of gamma GT than females. The association showed higher value of gamma GT for males compared to females in all the models, assuming other covariates in the models are fixed.

The results further suggested that there is no strong association between all the covariates and alkaline phosphate in all the models (p>0.05).

Table 7 is a summary result of analyses for total cholesterol, LDL, HDL and triglyceride outcomes with our selected PFAS compounds and patients' characteristics. The findings show that there is no association between any PFAS compound and health outcomes (p>0.05).

The models with PFOS (L+B), PFOA (L+B) and PFHxS (L+B) suggest that sex is a significant covariate for total cholesterol. Total cholesterol in males is 0,16 mg/dL (p=0.03) lower compared to females, holding the other factors constant. It also shows that smoking people have an increased total cholesterol compared to non-smoking (estimate: 0,20 mg/dL higher, p=0.04). No covariate is significantly associated with LDL in all the corresponding models. Males have significantly lower HDL than females. For example, in the model with PFOS (L+B), the amount of HDL in males is 0.18 mg/dL (p=0.01) lower than females, holding the other factors constant. A one kg/m^2 increase in BMI is associated with a significant decrease in HDL. There is no evidence for the significant association of any covariate with triglyceride. In conclusion, although some of the association between covariates and health effects were statistically significant in these models, the effects were very small and not clinically relevant.

Table 8 summarizes models based on continuous and binary outcomes. Age is significantly associated with a decrease in eGFR in all corresponding models. As the age of patients increased by 1 year, their eGFR decreased by 0.27 mL/min/1.73m^2 (based on the model with PFOS (L+B)) (p=0.03). There is a strong association between sex and TSH for the model with PFNA and with PFHxS (L). BMI is found to be statistically significantly associated with hypertension in all the models. Considering the model with PFOA (L) as corresponding exposure, as BMI increases by 1 kg/m^2 the odds of having a diagnosis of hypertension increase by 32% (OR=1.32, p=0.01). Age is not significantly associated with the status of hypertension. There is no association between any of the covariates in the model and arthrosis status.

Table 6: Estimates, standard error and p-values of multiple linear regression model for liver function parameters

PFAS compound	Covariates	ALT outcome			AST Outcome			Gamma G-T outcome			Alkaline Posphate outcome		
		β	SE	p	β	SE	p	β	SE	p	β	SE	p
PFOS (L)	PFOS	-0.02	0.04	0.60	0.02	0.03	0.65	-0.07	0.07	0.27	0.05	0.08	0.59
	Sex Male	2.86	2.72	0.30	-0.56	2.24	0.8	**11.6**	**4.50**	**0.01**	7.47	6.21	0.24
	Age	-0.16	0.12	0.18	0.02	0.09	0.86	0.05	0.19	0.79	0.01	0.24	0.98
	BMI	**0.57**	**0.27**	**0.04**	0.02	0.23	0.95	0.65	0.44	0.14	0.27	0.57	0.64
	Smoke Unkown	**-5.49**	**2.71**	**0.04**	-0.42	2.26	0.85	-3.87	4.50	0.39	-6.26	6.22	0.32
	Smoke Yes	-3.04	4.03	0.46	5.08	3.23	0.12	0.73	6.37	0.91	-7.47	9.02	0.41
PFOS (L+B)	PFOS	-0.02	0.04	0.63	-0.02	0.08	0.83	0.06	0.06	0.34	0.07	0.07	0.35
	Sex Male	2.72	3.23	0.40	-1.01	2.24	0.66	**11.7**	**4.43**	**0.01**	7.57	5.86	0.20
	Age	-0.16	0.14	0.25	0.01	0.09	0.91	0.03	0.19	0.88	-0.09	0.23	0.69
	BMI	0.57	0.32	0.08	0.17	0.21	0.41	0.65	1.51	0.14	0.28	0.55	0.61
	Smoke Unkown	-6.26	3.21	0.06	-1.03	2.16	0.64	-3.38	-0.77	0.45	0.16	5.87	0.97
	Smoke Yes	-0.24	4.59	0.96	3.46	3.07	0.27	-4.29	-0.68	0.50	1.69	8.29	0.84
PFOA (L)	PFOA	-0.84	1.06	0.43	-0.70	0.66	0.30	1.40	1.75	0.42	0.72	2.3	0.76
	Sex Male	1.72	2.81	0.54	-0.38	1.77	0.83	**9.79**	**4.68**	**0.04**	7.85	6.3	0.22
	Age	-0.09	0.13	0.48	0.07	0.08	0.42	-0.02	0.22	0.93	-0.11	0.29	0.71
	BMI	0.42	0.28	0.14	-0.02	0.17	0.91	0.58	0.46	0.21	0.51	0.59	0.39
	Smoke Unkown	-4.61	2.77	0.10	-1.40	1.73	0.42	-2.40	4.55	0.60	-1.97	6.11	0.75
	Smoke Yes	0.84	4.04	0.84	-0.56	2.52	0.82	-3.53	6.55	0.59	-4.3	8.76	0.63
PFOA (L+B)	PFOA	-0.04	0.04	0.28	-4.40	2.67	0.11	0.62	1.63	0.71	2.15	2.20	0.33
	Sex Male	0.05	0.11	0.68	-9.36	7.19	0.20	**11.4**	**4.48**	**0.013**	6.95	5.91	0.25
	Age	-0.01	0.01	0.84	**0.71**	**0.33**	**0.04**	0.01	0.21	0.99	-0.20	0.27	0.47
	BMI	**0.03**	**0.01**	**0.02**	-0.52	0.71	0.46	0.66	0.43	0.13	0.38	0.55	0.49
	Smoke Unkown	-0.15	0.11	0.31	5.32	7.03	0.45	-4.44	4.26	0.30	0.93	5.73	0.87
	Smoke Yes	-0.06	0.17	0.72	9.93	10.3	0.34	-6.25	6.13	0.31	0.32	8.49	0.97

PFAS compound	Covariates	ALT outcome			AST Outcome			Gamma G-T outcome			Alkaline Posphate outcome		
		β	SE	p	β	SE	p	β	SE	p	β	SE	p
PFHxS (L)	PFHxS	-0.48	0.5	0.34	-0.16	0.78	0.83	0.56	0.67	0.40	-0.02	1.16	0.98
	Sex Male	2.14	2.88	0.46	-5.27	4.37	0.24	**12.4**	**3.80**	0.002	9.32	5.52	0.10
	Age	-0.16	0.14	0.25	0.24	0.20	0.25	-0.02	0.18	0.90	-0.13	0.25	0.59
	BMI	**0.67**	**0.28**	0.02	-0.18	0.43	0.69	0.50	0.37	0.19	0.34	0.52	0.51
	Smoke Unkown	-5.08	2.79	0.08	1.84	4.23	0.67	-4.28	3.68	0.25	-4.21	5.43	0.44
	Smoke Yes	-0.67	4.06	0.87	4.19	6.22	0.51	-6.80	5.32	0.21	-3.20	7.76	0.68
PFHxS (L+B)	PFHxS	-0.55	0.56	0.33	-0.31	0.39	0.43	0.30	0.86	0.73	0.08	1.27	0.95
	Sex Male	0.68	3.13	0.83	-0.42	2.21	0.85	**13.5**	**4.88**	0.009	6.26	5.98	0.30
	Age	-0.10	0.15	0.49	0.10	0.10	0.36	0.04	0.23	0.86	-0.13	0.27	0.65
	BMI	0.56	0.31	0.08	0.16	0.22	0.46	0.67	0.48	0.17	0.30	0.57	0.60
	Smoke Unkown	-1.66	3.03	0.59	-0.70	2.15	0.74	-6.97	4.72	0.15	-2.71	5.89	0.64
	Smoke Yes	1.28	4.59	0.78	3.11	3.14	0.33	-521	6.82	0.45	-2.13	8.54	0.80
PFNA	PFNA	0.1	0.18	0.57	-2.22	2.47	0.38	2.62	5.51	0.64	0.72	2.3	0.76
	Sex Male	0.16	0.13	0.20	-1.06	1.74	0.55	**12.3**	**3.90**	0.001	7.85	6.3	0.22
	Age	-0.01	0.07	0.28	0.01	0.09	0.98	-0.03	0.20	0.89	-0.11	0.29	0.71
	BMI	**0.03**	**0.01**	0.04	0.01	0.18	0.99	0.68	0.40	0.10	0.51	0.59	0.39
	Smoke Unkown	-0.23	0.13	0.07	-0.40	1.7	0.82	-4.53	3.80	0.24	-1.97	6.11	0.75
	Smoke Yes	0.09	0.18	0.65	0.07	2.5	0.98	-6.70	5.52	0.23	-4.3	8.76	0.63

β: estimate; SE: standard error; p: probability

Table 7: Estimates, standard error and p-values of multiple linear regression model of cholesterol-related outcomes

PFAS compound	Covariates	Total cholesterol outcome			LDL outcome			HDL outcome			Triglyceride outcome		
		β	SE	p	β	SE	p	β	SE	p	β	SE	p
PFOS (L)	PFOS	-0.01	0.00	0.81	-0.01	0.01	0.95	-0.01	0.00	0.79	-0.01	0.00	0.76
	Sex Male	-0.1	0.07	0.15	-0.17	0.12	0.17	**-0.18**	**0.07**	**0.01**	0.07	0.14	0.60
	Age	-0.01	0.00	0.84	-0.01	0.01	0.41	0.00	0.00	0.37	0.01	0.01	0.25
	BMI	-0.01	0.01	0.27	-0.01	0.01	0.79	**-0.02**	**0.01**	**0.01**	0.01	0.02	0.67
	Smoke Unkown	-0.04	0.07	0.52	-0.12	0.13	0.35	0.01	0.08	0.91	-0.08	0.14	0.57
	Smoke Yes	0.09	0.1	0.38	0.16	0.18	0.37	-0.12	0.10	0.24	0.35	0.18	0.06
PFOS (L+B)	PFOS	0.01	0.00	0.88	0.01	0.01	0.89	0.01	0.00	0.95	-0.01	0.00	0.76
	Sex Male	**-0.16**	**0.07**	**0.03**	-0.16	0.12	0.19	**-0.18**	**0.07**	**0.02**	0.07	0.13	0.62
	Age	0.01	0.00	0.60	-0.01	0.01	0.82	0.01	0.0	0.4	0.01	0.01	0.20
	BMI	-0.01	0.00	0.16	-0.01	0.01	0.56	**-0.02**	**0.01**	**0.01**	0.02	0.01	0.14
	Smoke Unkown	-0.04	0.07	0.56	0.11	0.13	0.36	0.07	0.08	0.39	-0.27	0.14	0.07
	Smoke Yes	**0.20**	**0.10**	**0.04**	0.23	0.17	0.19	-0.02	0.1	0.83	0.09	0.18	0.62
PFOA (L)	PFOA	0.01	0.03	0.67	0.03	0.05	0.50	-0.01	0.03	0.64	-0.01	0.06	0.81
	Sex Male	-0.14	0.07	0.06	-0.13	0.13	0.32	**-0.24**	**0.07**	**0.00**	0.09	0.13	0.51
	Age	0.01	0.00	0.95	-0.01	0.01	0.39	0.01	0.01	0.16	0.01	0.01	0.25
	BMI	-0.01	0.00	0.14	-0.01	0.01	0.72	**-0.02**	**0.01**	**0.00**	0.01	0.02	0.59
	Smoke Unkown	-0.04	0.07	0.59	-0.03	0.13	0.84	0.04	0.07	0.58	-0.16	0.14	0.27
	Smoke Yes	0.24	0.10	0.03	0.24	0.18	0.19	-0.09	0.1	0.39	0.27	0.18	0.14
PFOA (L+B)	PFOA	-0.01	0.03	0.89	0.02	0.05	0.64	0.01	0.03	0.92	0.02	0.06	0.78
	Sex Male	**-0.16**	**0.08**	**0.04**	-0.09	0.13	0.49	**-0.19**	**0.07**	**0.01**	0.12	0.15	0.40
	Age	0.00	0.00	0.99	-0.01	0.01	0.41	0.00	0.00	0.97	0.01	0.01	0.51
	BMI	-0.01	0.01	0.26	-0.01	0.01	0.99	**-0.02**	**0.01**	**0.01**	0.02	0.02	0.16
	Smoke Unkown	-0.07	0.08	0.38	-0.10	0.13	0.45	0.05	0.07	0.50	-0.14	0.16	0.36
	Smoke Yes	0.15	0.10	0.16	0.09	0.17	0.61	-0.16	0.11	0.14	0.19	0.20	0.37

PFAS compound	Covariates	Total cholesterol outcome			LDL outcome			HDL outcome			Triglyceride outcome		
		β	SE	p	β	SE	p	β	SE	p	β	SE	p
PFHxS (L)	PFHxS	0.00	0.01	0.97	0.01	0.02	0.78	0.01	0.01	0.73	0.01	0.02	0.76
	Sex Male	-0.11	0.08	0.15	-0.18	0.12	0.13	**-0.23**	**0.07**	0.00	0.17	0.13	0.18
	Age	0.00	0.00	0.84	-0.01	0.01	0.82	0.00	0.00	0.31	0.00	0.01	0.59
	BMI	-0.01	0.01	0.32	-0.1	0.01	0.34	**-0.02**	**0.01**	0.00	0.01	0.01	0.33
	Smoke Unkown	-0.07	0.08	0.36	-0.07	0.12	0.55	0.02	0.07	0.80	-0.14	0.14	0.31
	Smoke Yes	0.15	0.11	0.18	0.24	0.17	0.17	-0.05	0.10	0.60	0.13	0.17	0.45
PFHxS (L+B)	PFHxS	-0.01	0.01	0.78	-0.01	0.02	0.58	0.01	0.01	0.39	-0.01	0.02	0.91
	Sex Male	**-0.18**	**0.07**	0.01	-0.09	0.12	0.46	**-0.20**	**0.07**	0.00	0.10	0.12	0.43
	Age	0.00	0.00	0.59	-0.02	0.00	0.67	-0.01	0.00	0.94	0.01	0.01	0.12
	BMI	-0.01	0.00	0.13	0.04	0.01	0.75	**-0.02**	**0.00**	0.02	0.02	0.01	0.08
	Smoke Unkown	-0.05	0.07	0.47	-0.09	0.12	0.46	0.06	0.07	0.41	-0.22	0.13	0.10
	Smoke Yes	0.19	0.1	0.08	0.22	0.16	0.17	**-0.21**	**0.10**	0.04	0.09	0.16	0.58
PFNA	PFNA	-0.03	0.10	0.73	-0.12	0.18	0.50	0.13	0.1	0.20	-0.02	0.20	0.94
	Sex Male	-0.14	0.07	0.05	-0.12	0.12	0.29	**-0.18**	**0.07**	0.01	0.10	0.14	0.45
	Age	0.00	0.00	0.99	-0.01	0.01	0.66	-0.01	0.01	0.84	0.01	0.01	0.38
	BMI	-0.01	0.00	0.22	0.00	0.01	0.97	**-0.02**	**0.01**	0.01	0.01	0.02	0.41
	Smoke Unkown	-0.08	0.07	0.25	-0.13	0.12	0.29	0.04	0.07	0.55	-0.22	0.14	0.13
	Smoke Yes	0.13	0.1	0.22	0.07	0.16	0.68	-0.09	0.1	0.36	0.26	0.19	0.17

β: estimate; SE: standard error; p: probability

Table 8: Estimates, standard error and p-value based on multiple linear regression and logistic regression for eGFR, TSH, hypertension and arthrosis

PFAS compound	Covariates	Continous Outcomes						Binary Outcomes					
		eGFR			TSH			Hypertension			Arthrosis		
		β	SE	p	β	SE	p	Exp(β)	SE	p	Exp(β)	SE	p
PFOS (L)	PFOS	-0.07	0.04	0.05	-0.01	0.00	0.95	1.00	0.01	0.83	1.00	0.01	0.85
	Sex Male	2.45	2.60	0.35	-0.32	0.36	0.38	2.82	0.79	0.19	0.46	0.73	0.29
	Age	**-0.46**	**0.11**	**0.00**	0.02	0.01	0.27	1.04	0.03	0.31	1.04	0.03	0.31
	BMI	-0.01	0.28	0.97	0.03	0.03	0.46	**1.28**	**0.09**	**0.01**	0.99	0.08	0.94
	Smoke Unkown	-1.87	2.58	0.47	0.04	0.36	0.91	1.73	0.84	0.51	0.54	0.77	0.43
	Smoke Yes	-1.70	3.47	0.63	-0.78	0.50	0.13	0.64	1.04	0.67	1.19	0.95	0.86
PFOS (L+B)	PFOS	-0.06	0.04	0.09	0.00	0.00	0.91	1.01	0.01	0.80	1.00	0.01	0.83
	Sex Male	0.62	2.87	0.83	-0.51	0.39	0.20	2.81	0.79	0.19	0.46	0.73	0.29
	Age	**-0.27**	**0.12**	**0.03**	0.02	0.02	0.24	1.04	0.04	0.32	1.04	0.03	0.31
	BMI	0.27	0.29	0.37	0.03	0.04	0.51	**1.28**	**0.09**	**0.01**	0.99	0.08	0.94
	Smoke Unkown	0.06	2.90	0.98	0.36	0.41	0.38	1.74	0.84	0.51	0.54	0.77	0.43
	Smoke Yes	1.63	4.09	0.69	-0.34	0.58	0.57	1.65	0.04	0.68	0.18	0.95	0.86
PFOA (L)	PFOA	-0.97	1.03	0.35	0.16	0.12	0.19	1.57	0.31	0.14	0.77	0.31	0.40
	Sex Male	2.10	2.91	0.47	-0.58	0.29	0.06	2.82	0.81	0.20	0.47	0.74	0.30
	Age	**-0.36**	**0.13**	**0.01**	0.01	0.01	0.46	1.01	0.04	0.79	1.05	0.04	0.21
	BMI	-0.12	0.31	0.71	0.01	0.02	0.75	**1.32**	**0.10**	**0.00**	0.99	0.08	0.89
	Smoke Unkown	1.29	2.86	0.65	0.24	0.31	0.43	2.16	0.82	0.35	0.50	0.76	0.36
	Smoke Yes	2.20	4.05	0.59	-0.17	0.39	0.66	0.82	1.07	0.86	1.04	0.96	0.97
PFOA (L+B)	PFOA	-1.18	0.89	0.19	0.05	0.13	0.72	1.55	0.31	0.16	0.73	0.32	0.33
	Sex Male	1.18	2.56	0.65	-0.43	0.32	0.18	2.94	0.81	0.18	0.56	0.75	0.43
	Age	**-0.29**	**0.12**	**0.02**	0.01	0.01	0.33	1.01	0.04	0.79	1.05	0.04	0.21
	BMI	0.23	0.27	0.40	0.03	0.03	0.31	**1.30**	**0.10**	**0.01**	0.94	0.09	0.50
	Smoke Unkown	1.39	2.51	0.58	0.33	0.32	0.32	2.26	0.83	0.32	0.60	0.78	0.51
	Smoke Yes	-0.31	3.46	0.93	-0.15	0.42	0.72	0.86	1.06	0.89	1.19	0.97	0.86

PFAS compound	Covariates	Continous Outcomes						Binary Outcomes					
		eGFR			TSH			Hypertension			Arthrosis		
		β	SE	p	β	SE	p	Exp(β)	SE	p	Exp(β)	SE	p
PFHxS (L)	PFHxS	-0.30	0.53	0.57	0.03	0.05	0.59	0.89	0.14	0.40	0.87	0.16	0.36
	Sex Male	1.64	3.10	0.60	**-0.63**	**0.27**	0.03	2.84	0.80	0.19	0.46	0.74	0.29
	Age	**-0.36**	**0.15**	0.02	0.01	0.01	0.45	1.06	0.04	0.20	1.05	0.04	0.21
	BMI	-0.05	0.32	0.87	0.02	0.02	0.43	**1.29**	**0.09**	0.00	1.00	0.08	0.95
	Smoke Unkown	0.38	2.98	0.90	0.22	0.28	0.43	1.44	0.80	0.65	0.51	0.76	0.37
	Smoke Yes	1.70	4.24	0.69	-0.17	0.35	0.63	0.54	1.03	0.55	1.16	0.94	0.87
PFHxS (L+B)	PFHxS	-0.22	0.45	0.63	0.04	0.07	0.55	0.88	0.14	0.35	0.81	0.19	0.26
	Sex Male	1.54	2.57	0.55	-0.54	0.35	0.14	3.08	0.80	0.16	0.55	0.76	0.42
	Age	**-0.42**	**0.12**	0.00	0.01	0.02	0.64	1.06	0.04	0.19	1.06	0.04	0.18
	BMI	-0.08	0.28	0.77	0.02	0.03	0.46	**1.26**	**0.09**	0.01	0.96	0.09	0.62
	Smoke Unkown	1.90	0.44	0.44	-0.03	0.36	0.93	1.58	0.81	0.57	0.63	0.78	0.55
	Smoke Yes	1.49	3.54	0.68	-0.43	0.47	0.36	0.60	1.02	0.61	1.41	0.95	0.72
PFNA	PFNA	-5.90	3.60	0.11	0.21	0.40	0.60	2.31	1.06	0.43	0.50	1.15	0.54
	Sex Male	1.94	2.67	0.47	**-0.59**	**0.28**	0.04	3.04	0.80	0.16	0.53	0.75	0.39
	Age	**-0.30**	**0.14**	0.04	0.01	0.01	0.42	1.02	0.04	0.69	1.05	0.04	0.26
	BMI	-0.22	0.30	0.48	0.02	0.03	0.34	**1.29**	**0.10**	0.01	0.93	0.09	0.45
	Smoke Unkown	0.97	2.60	0.71	0.26	0.28	0.36	1.80	0.79	0.46	0.69	0.76	0.62
	Smoke Yes	1.71	3.69	0.65	-0.14	0.36	0.71	0.61	1.02	0.63	1.52	0.95	0.66

β: estimate; SE: standard error; p: probability

5.2.1 Mixture analysis

We extended the separate model analysis to mixture analysis in order to investigate the mixture effect of various PFAS compounds on health outcomes. There is a perfect positive correlation between linear and linear+branched of the same pollutant (for example PFOA (L) and PFOA (L+B), PFOS (L) and PFOS (L+B)). As a result, we selected the linear PFAS values to include in the mixture analysis. Table 9 summarizes the weighted quantile sum (WQS) regression models for the mixtures analysis including (PFOS (L), PFOA (L),_PFHxS (L), PFNA). The results suggested that there is no statistically significant association between exposure to PFAS mixtures and ALT or hypertension outcomes.

Table 9: Summary results of the WQS regression for linear and binary outcomes

Covariates	Continuous outcome (ALT)			Binary outcome (hypertension)		
	β	SE	p	β	SE	p
Wqs	-0.42	1.23	0.74	0.07	0.09	0.45
Age	0.09	0.11	0.44	0.01	0.01	0.86
Sex Male	3.43	2.48	0.18	0.10	0.19	0.61
BMI	0.28	0.27	0.30	0.03	0.028	0.22
Smoke Unknown	-1.92	2.52	0.45	-0.09	0.19	0.65
Smoke Yes	-0.74	3.51	0.83	-0.28	0.28	0.33

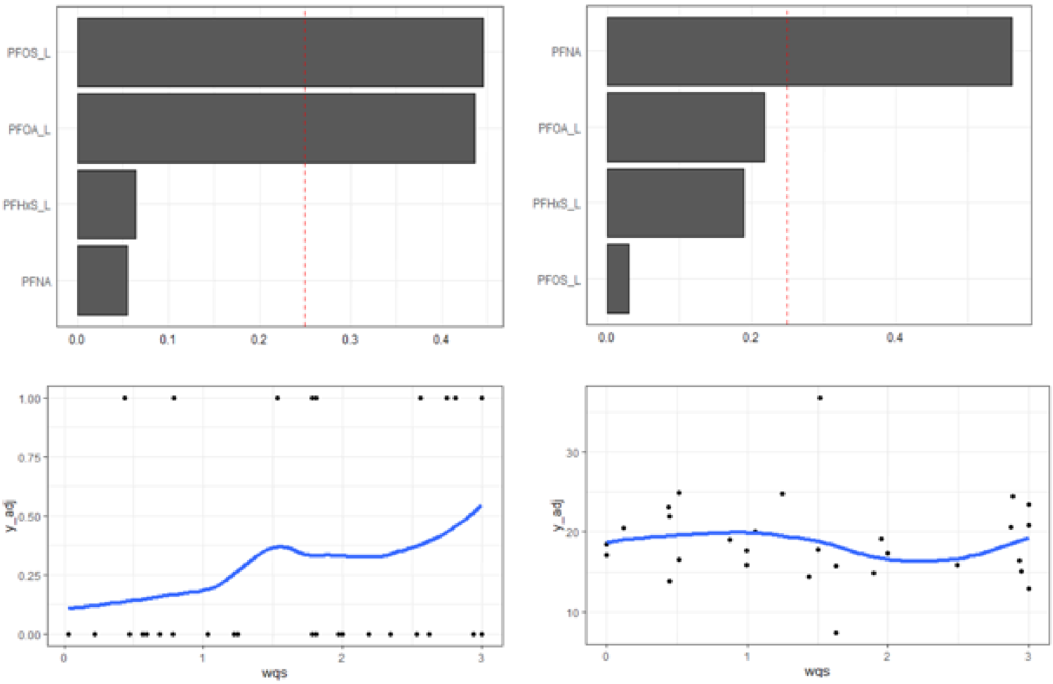

Figure 3: Bar plots for weights of PFAS (top-left: ALT, top-right: hypertension) and scatter plot for the association between weights and outcomes (bottom-left: ALT, bottom-right: hypertension)

The bar plots in Figure 3 (top left and top right) show the weights assigned to each PFAS compound ordered from the highest weight to the lowest. For the ALT outcome, PFOA (L) and PFOS (L) seem the compounds with the highest weight that contributed to the mixture though we confirmed that the association between mixture exposure and outcomes is not significant based on Table 9. For hypertension, on the other hand, PFNA contributed relatively the highest weight. The scatter plot in Figure 3 (bottom left and bottom right) represents the WQS index versus outcome after adjusting for the covariates. It shows the direction and shape of the association. The plots show that there is no clear indication of either a positive or negative association between the PFAS and outcomes, except for a slight increase for ALT, which is in support of the finding in Table 9.

Table 10 represents the overall mixture effect on ALT and hypertension. The result suggest that the mixture effect is not statistically significant for any outcome which is in accordance with the result obtained based on weighted quantile regression. The result is obtained after adjusting for covariates. Figure 4 explores the weight of each PFAS and the direction of their contribution to the two outcomes. The direction of the contribution of some PFAS types is not consistent between the two outcomes. However, as the overall mixture effect is not significantly associated with the outcomes, the direction of contribution and their magnitude is not important. The overall mixture effect on ALT and hypertension is not statistically significant.

Table 10: The overall mixture effect based on g-computation

Outcomes	Parameter	Estimate	Standard error	p-value
ALT	Psi1	-1.15	1.72	0.51
Hypertension	Psi1	0.54	0.49	0.27

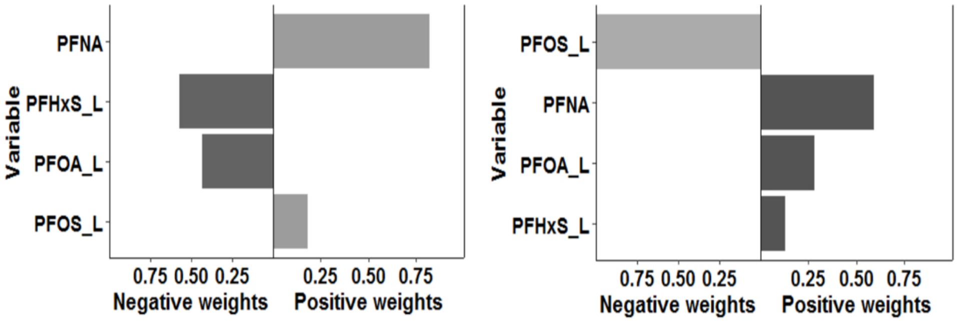

Figure 4: Weights based on g-computation (ALT: left, Hypertension: right)

Table 11 shows that the posterior inclusion probability (PIP) for all PFAS is not approximate to 1, suggesting that there is no strong evidence to include any of them in the model. Figure 5 also illustrates the exposure-response relationship that an increase in one of the PFAS types is not strongly associated with either an increase or a decrease in outcomes (ALT or hypertension) for all other PFAS types at the 50th percentile, while controlling for covariates.

Table 11: Posterior inclusion probability estimate

Mixture PFAS	Outcome = ALT	Outcome = hypertension
	PIP	PIP
PFOA-linear	0.19	0.57
PFOS-linear	0.27	0.51
PFNA	0.43	0.38
PFHxS-linear	0.18	0.42

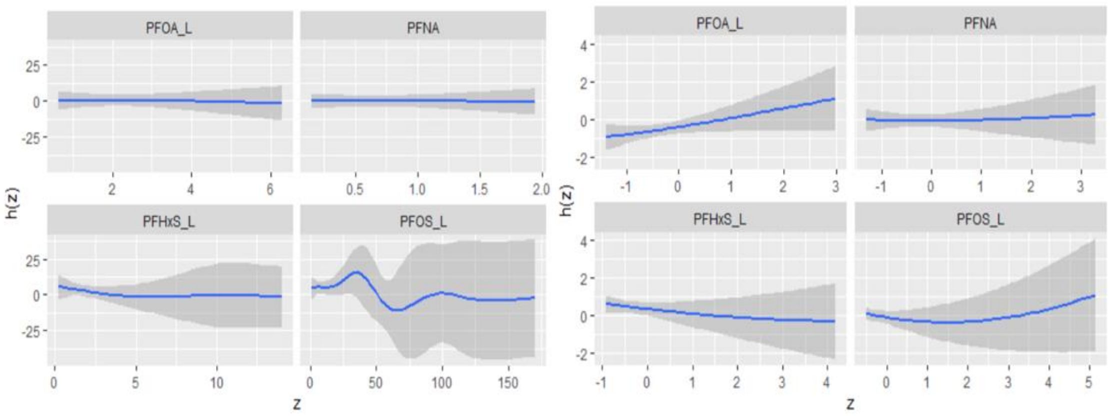

Figure 5: Predictor-response function (Left: ALT, right: Hypertension)

CHAPTER 6: DISCUSSION

6.1 GENERAL DISCUSSION

In this feasibility study, we wanted to study whether it is possible to investigate associations between exposure to chemicals (in this case PFAS) and health outcomes, by combining exposure data and health data originating from different dataset, i.e. by combining serum PFAS levels obtained in human biomonitoring studies with the primary health care data from the Flemish Intego database. The goal of this feasibility study was not to determine whether PFAS compounds have health effects, since the amount of data is too small to make sound conclusions. However, the foundation created by this feasibility study can be used to do future analyses on larger amounts of data, so associations between chemical exposures and health outcomes can be studied using primary care health data.

In a first step, a database was created with the PFAS values coupled to patient and GP identifiers. This database was then transformed into .lab and .adr-files using a macro. Next, these files with coded PFAS values were successfully sent to the medical file of GPs using the UM module. To ensure that PFAS values arrived in the correct EMR, coupling to the social security number was necessary. Coded data in the EMR of GPs participating in Intego is extracted to the HealthData environment on a weekly basis. This coded data includes lab data such as PFAS values, but also numerous health outcomes or demographic information. The data present in the HealthData environment can then be transferred to the Intego environment, where analyses are possible. Having both the exposure and outcome data in the Intego database at the level of the patient (‘coupled health and exposure data’) made it possible to perform a limited number of analyses. Where we thought only descriptive analyses would be possible, some statistical analyses were possible as well, despite the relatively small amount of data.

The models we used incorporated several covariates for every outcome. The fact that the effects of these covariates are in the direction we would expect them to be (e.g. decreasing kidney function with increasing age) is an internal validation of these models. This feasibility study proves that we can analyse different outcome types (binary, continuous,…) with our technique on the coupled data.

Our results show no significant impact of PFAS on any of the health effects investigated. However, as mentioned, the amount of data was too small to make a conclusion on health effects of PFAS. The important take away message from this project is the feasibility of this method and our ability to use the created techniques and scripts to perform analyses on larger datasets. With more exposure data (i.e. PFAS levels in serum) becoming available in the ongoing large-scale monitoring study (https://www.vlaanderen.be/pfas-vervuiling/pfas-bloedonderzoeken-algemeen) we will be able to do more robust analyses and make valid conclusions, at least for some health effects.

6.2 STRENGTHS

Our study design has several important strengths.

It is the first time that we are able to combine chemical exposure and health data originating from different datasets (originally designed for different purposes) into a combined dataset including chemical exposure and health data at the detailed level of individuals in a register. Usually, when investigating exposure-response associations in research projects, paired data are collected in the participants by measuring biomarkers of exposure in combination with biomarkers of effect (either measured or assessed through questionnaires). Gathering such data is labour intensive, both for the participant and the research team. We now can supplement (or partly replace) HBM studies by using health data of participants that are routinely collected for primary care purposes. This also means that we can investigate associations with a wide range of effects, typically beyond the number of effects that we can investigate using questionnaires and biomarker of effects in HBM studies.

The collection of both exposure and outcome data in one database allows to perform the analyses of exposure-response associations in an efficient manner. The amount of health data and demographic data that are available in the Intego database creates an opportunity to analyse many different health effects available in primary care settings. Further, once a model for a certain outcome is created (outcome definition, statistical plan,…), it can be used in future studies. In other words, much of the work regarding study design can be used in the same way in other studies linking chemical exposure and health effects, provide that adequate number of patients with paired exposure and health data can be generated. Finally, since health data is extracted to Intego on a regular basis, analyses can be carried out at certain time intervals to find time trends in PFAS exposure and health effects (e.g. effects that are only identified after a certain lag time). Repeated analyses can provide additional information, supplementary to cross-sectional studies.

6.3 LIMITATIONS & SOLUTIONS

Every study has its limitations to overcome.

The Intego database cannot distinguish which PFAS result was analysed in which lab. It is possible that some serum PFAS measurements were requested by individual GPs or third-party organizations and were analysed in non-accredited labs. We expect this number to be small however, since these individual initiatives are small in comparison with the large-scale studies where a few hundred, up to a few thousands of measurements are performed in accredited labs. Creating guidelines and recommendations for future HBM-studies in Flanders can overcome this problem, cfr. Chapter 7.

Furthermore, it is important to note that the outcomes used originate from the EMR of GPs. It is possible that these outcomes are an underrepresentation of the actual prevalence or incidence in the general population. Not all people with a cold, for example, will visit their GP and have this diagnosis coded into their file. Also, people might have an increased blood pressure at home, without a diagnosis of hypertension in their file. However, since the GPs participating in Intego have to adhere to certain quality criteria, this underrepresentation is minimalized and most likely even smaller than studies using for example patient-reported outcomes.

When investigating outcomes with a low prevalence or incidence, two problems can occur. If we investigate a rare disease in a geographically defined area, identification of patients may become possible and thus analysis cannot be carried out because of conflicts with privacy issues. Second, our analyses will not have enough power if the outcome counts are too low. To overcome this problem, we can broaden the time interval in which we measure the outcome (e.g. more myocardial infarctions in 5 years than in 1 year); group together different outcomes (cfr. 4.1), or enlarge the dataset (this might prevail in the future when the results of the large-scale monitoring study are received in the EMR of the GP).

The exposure should be quantifiable in at least a substantial number of patients or study participants. Guidance on minimal LOD or LOQ when analysing chemicals in serum or urine can minimize the number of non-quantifiable samples. We can use imputation techniques if only a small portion of the values is below the LOQ. If a higher number of values is below the LOQ, we can transform them into a binary exposure and compare health outcomes in the group of participants with values above the LOQ to the group of participants with values below the LOQ.

Lastly, some data simply isn’t available in Intego. For example, data on hospitalization or mortality can be better studied using other databases. Linking these other databases with Intego can create an even more complete picture of pollution-related health problems.

CHAPTER 7: RECOMMENDATIONS

In conclusion, we have investigated the possibility to enrich the Intego database with chemical exposure data (serum PFAS values) from external biomonitoring studies and using this enriched database to investigate associations between chemical exposure and health effects. This feasibility study shows that it is indeed possible to do this. With a larger amount of data than included in the current feasibility study database, we will be able to make more robust analyses and determine health effects of PFAS in primary care with more certainty.

We want to formulate some recommendations for future studies to improve upon our methodology.

7.1 COLLECTION AND TRANSFER OF DATA

The current proof-of-concept analysis shows that linking HBM data to health records of primary care is a promising option for future research. Therefore, it is important that some essential aspects are included in the design of HBM studies, to pro-actively allow the transfer of HBM data to the EMR of the GP.

It is recommended to ask consent from the participant to send the results to the GP in the informed consent form. It is important to stress that transferring the individual HBM data to the GP is in the mutual benefit of both the participant (since these exposure data might be of importance for the individual follow-up in the future) and for the public good, since this allows more and better scientific research in the future.

In order to allow a consistent and high-quality analysis of the data with a minimum of noise, it is essential that the toxicological data are of high quality. Therefore, in the planning of HBM studies, it is recommended to select accredited laboratories only. Also, if several studies are performed by different labs, accreditation guarantees that the results are comparable and can be pooled to allow a robust statistical analysis.

The above recommendations can be given both for the commissioning bodies that initiate the study (and define the boundary conditions) and to the researchers that perform the study.

Additionally to the quality of the measurement itself, it is also important that the terminology and nomenclature is standardized. This is a commitment that should come from the providers of software packages for EMRs. At the moment, every software package uses its own coding rules. However, there is a need for a uniform nomenclature and standardized coding rules. Currently, a new standard format is developed, namely ‘FHIR’. From a research point of view, we recommend to promote this new standard.

7.2 STATISTICAL ANALYSIS

Some limitations that are present, may be overcome by taking extra measures or can be tackled in the statistical analysis.

The health effects that are extracted from the EMRs will be mainly diagnoses of diseases, which may lead to an underestimation of the true prevalence or incidence of the diseases of interest. This would not be the case if all patients were examined for the disease of interest, and is neither the case for health parameters that are not by definition linked to a diagnosis of a disease and are widely monitored, e.g. blood pressure. Also, health data can be biased: some patients will be continuously followed up for specific biomarkers by the GP as they have a certain pathology, whilst others will not have any value for the biomarker in their EMR. These aspects should be taken into account in the interpretation of the statistical results and in the statistical analysis, e.g. by performing sensitivity analysis on subgroups of patients with more complete data or with more follow-up data. Also, a thorough exploration of the data will help to assess the strengths and weaknesses. E.g. internal validation can be done by interpreting the covariates: if effects are assessed in the direction and magnitude we expect, this is an internal validation of the data.

When using cross-sectional data (from one timepoint), no conclusions can be made about the time window of exposure and its relationship with the health effect. In other words, there is no way to determine whether exposure or health effect came first, which makes it harder to confer causality from an association. Also, for some health effects, there might be a lag time between exposure and onset of effects. These aspects should also be taken into account in the interpretation of the data. From a statistical point of view, analysis on different time points, with possible increasing effects over time can help to build strong evidence for causality.

The current study has already gathered important information on the type of health outcomes that are relevant, has constructed case definitions and has designed statistical scripts in R, both for single pollutant models and for mixed models. These methodologies can be made available for use in future studies. If new data are available, these can be analyzed with the current knowledge, and afterwards all steps of the methodology can be improved and refined further.

7.3 BEYOND INTEGO

In this study, we have investigated the possibility to enrich the Intego database with chemical exposure data (e.g. PFAS) originating from external studies, and using this enriched database to investigate associations between chemical exposure and health effects. In principle, this method could be used also to enrich other health databases (e.g. hospital registers, IMA register) with chemical exposure data and study associations between chemical exposure and health using other datasets. For each health database, some hurdles will be needed to overcome (e.g. technical aspects, GDPR, etc).

REFERENCES

ATSDR, 2021. Agency for Toxic Substances and Disease RegistryToxicological Profile for Perfluoroalkyls. Agency for Toxic Substances and Disease Registry. Atsdr 24.

Canova, C., Barbieri, G., Zare Jeddi, M., Gion, M., Fabricio, A., Daprà, F., Russo, F., Fletcher, T., Pitter, G., 2020. Associations between perfluoroalkyl substances and lipid profile in a highly exposed young adult population in the Veneto Region. Environ. Int. 145. https://doi.org/10.1016/j.envint.2020.106117

Catelan, D., Biggeri, A., Russo, F., Gregori, D., Pitter, G., Da Re, F., Fletcher, T., Canova, C., 2021. Exposure to perfluoroalkyl substances and mortality for covid-19: A spatial ecological analysis in the veneto region (italy). Int. J. Environ. Res. Public Health 18, 1–12. https://doi.org/10.3390/ijerph18052734

CPCSSN Team, Case Definitions: Canadian Primary Care Sentinel Surveillance Network (CPCSSN), Version 2022-Q4. February 6, 2023. https://cpcssn.ca/wp-content/uploads/2023/03/CPCSSN-Case-Definitions-2022-Q4_v2.pdf (*consulted on 15/01/2024*).

Delvaux, N., Aertgeerts, B., Van Bussel, JCH., Goderis, G., Vaes, B., Vermandere, M., 2018. Health data for research through a nationwide privacy-proof system in Belgium: design and implementation. JMIR Med. Inform. 6(4): e11428. doi: 10.2196/11428.

Gallo, E., Barbiellini Amidei, C., Barbieri, G., Fabricio, A.S.C., Gion, M., Pitter, G., Daprà, F., Russo, F., Gregori, D., Fletcher, T., Canova, C., 2022. Perfluoroalkyl substances and thyroid stimulating hormone levels in a highly exposed population in the Veneto Region. Environ. Res. 203. https://doi.org/10.1016/j.envres.2021.111794

Hammarstrand, S., Jakobsson, K., Andersson, E., Xu, Y., Li, Y., Olovsson, M., Andersson, E.M., 2021. Perfluoroalkyl substances (PFAS) in drinking water and risk for polycystic ovarian syndrome, uterine leiomyoma, and endometriosis: A Swedish cohort study. Environ. Int. 157, 106819. https://doi.org/10.1016/j.envint.2021.106819

Ottenbros, I., Govarts, E., Lebret, E., Vermeulen, R., Schoeters, G., Vlaanderen, J., 2021. Network Analysis to Identify Communities Among Multiple Exposure Biomarkers Measured at Birth in Three Flemish General Population Samples. Front. Public Heal. 9, 1–10. https://doi.org/10.3389/fpubh.2021.590038

Pitter, G., Zare Jeddi, M., Barbieri, G., Gion, M., Fabricio, A.S.C., Daprà, F., Russo, F., Fletcher, T., Canova, C., 2020. Perfluoroalkyl substances are associated with elevated blood pressure and hypertension in highly exposed young adults. Environ. Heal. A Glob. Access Sci. Source 19, 1–11. https://doi.org/10.1186/s12940-020-00656-0

Schulz, K., Silva, M.R., Klaper, R., 2020. Distribution and effects of branched versus linear isomers of PFOA, PFOS, and PFHxS: A review of recent literature. Sci. Total Environ. 733, 139186. https://doi.org/10.1016/j.scitotenv.2020.139186

VITO, PIH, 2021. bevolkingsonderzoek PFAS bij omwonenden van de 3M site in Zwijndrecht. Technisch wetenschappelijk rapport.

Xu, Y., Jurkovic-Mlakar, S., Lindh, C.H., Scott, K., Fletcher, T., Jakobsson, K., Engström, K., 2020a. Associations between serum concentrations of perfluoroalkyl substances and DNA methylation in women exposed through drinking water: A pilot study in Ronneby, Sweden. Environ. Int. 145. https://doi.org/10.1016/j.envint.2020.106148

Xu, Y., Li, Y., Scott, K., Lindh, C.H., Jakobsson, K., Fletcher, T., Ohlsson, B., Andersson, E.M., 2020b. Inflammatory bowel disease and biomarkers of gut inflammation and permeability in a community with high exposure to perfluoroalkyl substances through drinking water. Environ. Res. 181, 108923. https://doi.org/10.1016/j.envres.2019.108923