HEALTHYCLOUD
Health Research & Innovation Cloud

D7.1 Functional requirement analysis report of Use Case 1

Document Information

Contract Number	965345
Project Website	http://www.healthycloud.eu/
Contractual Deadline	M15, May 2022
Dissemination Level	PU
Nature	R
Author(s)	Emilie Cauët Marc Van Den Bulcke Irène Kesisoglou
Contributor(s)	
Reviewer(s)	Alicia Martínez-García (SAS) Michaela Mayrhofer (BBMRI-ERIC) Celia Álvarez (SAS) Helena Lodenius (CSC)
Keywords	Cancer, polygenic risk score, gene-environment interaction

Change Log

Version	Author	Date	Description of Change
V0.1	Juan González-García	08/03/2022	Initial Draft
V0.2	Emilie Cauët	19/04/2022	Draft submitted to reviewers
V0.3	Emilie Cauët	25/05/2022	Updated with review requests
V1.0	Juan González-García	27/05/2022	Styles adjust to be submitted
			(Final Change Log entries reserved for releases to the EC)

Table of contents

Executive Summary

The aim of WP7 is to demonstrate the utility of the envisioned HealthyCloud infrastructure by analysing real-world use cases for which specific research questions are addressed. The use case 1, led by Sciensano, focusses on cancer condition. It aims to assess how genomic information, gathered at the population level, can contribute to developing high-risk profiling for the major risk factors for cancer and other factors such as socio-economic status and health literacy. The deliverable 7.1 describes the necessary data for the use case including the existing Belgian data and the potential data from other countries (i.e., Finland, Spain and Germany) that may be included in future revisions. The problems encountered to access these data are described. A broad introduction to the statistical analysis for studying the effect of exposures to risk factors on the cancer outcome is provided. The functional requirements are expressed in terms of the analysis elements, addressed in Section 3, that will be latter included as inputs in Task 7.3 in conjunction with the functional requirements of D7.2.

1. Background

1.1. Use case description

The cancer use case 1, implemented in work package 7 (entitle 'Reference use cases as mechanisms to evaluate specifications') of the HealthyCloud project entails a feasibility study to assess how genomic information, gathered at the population level, can contribute to developing high-risk profiling for the major cancer risk factors (e.g. tobacco, alcohol, sun-exposure, …) and other factors such as socio-economic status and health literacy[1-3]. Combining genomic information with lifestyle-related factors and environmental exposure that affect cancer risk will allow us to investigate whether the association between a polygenic risk score (PRS), which in the context of the use case 1 reflects the cumulative effect of previously identified cancer-related genetic variants is modified by environmental factors. These findings could lead to an increased understanding of gene-environment interactions underlying the etiology of certain types of cancer which remains relatively unexplored in the cancer literature.

To that end, the cancer use case will gather information from different data collections. The exercise will initially be developed using Belgian and Finnish data in which the data collections/registries are national data collections and hence can act as a central point of contact. As there are many countries in Europe that organize their health data in a different (i.e. de-centralized) way, the acquired methodology could be expanded at a later stage in order to involve populations from other countries, such as Spain or Germany.

1.2. Use case opportunities

One of the final objectives of the cancer use case is to establish a generic workflow to combine various types of data for more understanding of the interplay between genetics and lifestyle and environmental factors in the origins of cancer disease. This generic workflow could be tested with cancers selected for study as proof-of-concepts.

The different steps are:

i) Explore national and/or regional data collections available (in Belgium and other countries);

---

[1] Khoury, M. J.; Holt, K. E. The Impact of Genomics on Precision Public Health: Beyond the Pandemic. *Genome Med* **2021**, *13* (1), 67, s13073-021-00886-y. https://doi.org/10.1186/s13073-021-00886-y.

[2] Koehly, L. M.; Persky, S.; Philip Shaw; Bonham, V. L.; Marcum, C. S.; Sudre, G. P.; Lea, D. E.; Davis, S. K. Social and Behavioral Science at the Forefront of Genomics: Discovery, Translation, and Health Equity. *Social Science & Medicine* **2021**, *271*, 112450. https://doi.org/10.1016/j.socscimed.2019.112450.

[3] Morris, T. T.; Davies, N. M.; Hemani, G.; Smith, G. D. Population Phenomena Inflate Genetic Associations of Complex Social Traits. *Sci. Adv.* **2020**, *6* (16), eaay0328. https://doi.org/10.1126/sciadv.aay0328.

ii) Develop the database structure amenable to analyse the combined genomic-lifestyle/environment information;

iii) Explore and develop statistical approaches (i.e. polygenic risk score methods and gene-environment interactions) to translate the integrated genomic information into cancer-risk questions;

iv) Evaluate the logs (issues, protocol deviations, lessons learned) and data sources and registers (risk, quality);

v) Define improvements (data harmonization, federated learning, cross-validation).

All these steps will help us to better understand the challenges of the whole HealthyCloud project and to demonstrate the utility of its envisioned infrastructure.

1.3. Use case challenges

For the cancer use case, we are endeavoring to link individual-level data from different sources such as genetics data and lifestyle factors in order to increase our understanding about the disease risk.

It will thus be necessary to take into consideration:

1) a variety of data collections required to respond to the research question;

2) the collection of structured and/or unstructured individual level data,

3) a subset of surveys and lifestyle questions that will need to be translated into trait names,

4) data with either population-based/cohort or non-population-based information,

5) data collected with different time spans and with different update periodicity.

2. Data requirements

2.1. Existing data

For Belgium, the cancer use case will gather information from the National Health Interview and Health Examination Survey 2018 (BELHISHES)[4], regarded as the principal reference in terms of population-based health statistics in Belgium. These data, used as a pilot case, will be linked to the Belgian Cancer Register (BCR) (cancer data) and Statbel (socio-economic data) (for more details, see annexes I and II).

[4] Nguyen, D.; Hautekiet, P.; Berete, F.; Braekman, E.; Charafeddine, R.; Demarest, S.; Drieskens, S.; Gisle, L.; Hermans, L.; Tafforeau, J.; Van der Heyden, J. The Belgian Health Examination Survey: Objectives, Design and Methods. *Arch Public Health* **2020**, *78* (1), 50. https://doi.org/10.1186/s13690-020-00428-9.

- **Belgian Health Interview and Health Examination Survey (BELHISHES) 2018**

The National Health Interview and Health Examination Survey 2018 (BELHISHES) is designed to obtain information on:

- People's health experience;
- Their health-related attitudes and behaviors;
- The extent to which they use health care facilities and;
- Their use of preventive health and social services

The participants of the BELHIS18 and BELHES18 have directly been surveyed and biological and physical determinations have been performed (i.e. individual data are not anonymous in the collection process). In particular, whole-genome sequencing (WGS) of the BELHES18 participants has been performed. In the cancer use case, a subset of survey data (lifestyle and environmental factors, chronic diseases profiles) associated with biological samples will be compiled (for more details, see annex I).

- **Belgian Cancer Registry (BCR)**

The Belgian Cancer Registry (BCR) is a core component of cancer control strategy in Belgium. BCR has a legal basis to collect data on all cases of cancer occurring in Belgian residents. The parameters coming from the BCR that will be included are the following:

- Cancer type;
- Date of first microscopic (cytological/histological) confirmation of malignancy or date of clinical/technical diagnosis if no microscopic examination was performed;
- Extent of disease.

- **Statbel (socio-economic data)**

National Population Register from Statistics Belgium (StatBel) provides information on the size of the population and demographic characteristics of the Belgian population. It also contains individual-level socio-economic data, for example the household composition of every citizen.

2.2. Desired data

Referring back to the general objectives of the use case study, if successful, the acquired methodology could be extended to populations from other countries if data sets are available.

Because these data come from different studies, they will need to be harmonized according to a common data dictionary.

The data listed below are the Finnish data collections that we may use to answer the use case on cancer:

- *FinHealth 2017 survey, Health Examination Survey*
- *Finnish Cancer Registry*
- Sosiaalihuollon hoitoilmoitusrekisteri (The Care Register for Social Welfare)
- Research Services at Statistics Finland
- Finnish Social Science Data Archive
- Avohilmo, Register of Primary Health Care Visits
- THL Biobank
- Findata
- FinSote, Health Interview Survey

As mentioned above, other countries, such as Germany and Spain, may be included at a later time point (section 2.3).

In terms of other types of data, imaging data such as data sets available through Euro-Bio-imaging ERIC or individual data on chemical exposures coming from bio-monitoring initiatives could also be considered for the use case and would be of great interest for this kind of study.

2.3. Data access challenges

The data collections/registries required to answer the cancer use case research question have been explored in conjunction with WP3.

At first, it is important to note that only Finland has a common descriptive metadata catalogue where researchers can find the available data collections in Finland and information on how to access the data from these data collections (for more details, see annexes III and IV). In Belgium, Spain or Germany there is no common national metadata catalogue yet, where a researcher could have an overview of all available data collections and the way to access data from those.

The data collections/registries in Belgium and Finland are national data collections and hence can act as a central point of contact, providing data from the whole country. On the contrary, in Spain and Germany the health data management system is very fragmented and data are stored in regional data collections and registries. Therefore, although we were able to contact and collect information on the data collections required in Finland and Belgium, it has been more challenging for Spain and Germany.

In Belgium and Finland we were able to contact individually the different data collections presented in annex (for more details, see annexes II and III). We gathered information on their data governance, the nature and granularity of the data they store and control, the quality of the datasets and the compliance of the data collection with the FAIR principles (findability, accessibility, interoperability and re-usability of the data and metadata).

Country	Data infrastructures
Finland	FinHealth 2017 survey, Health Examination Survey
Finland	Sosiaalihuollon hoitoilmoitusrekisteri (The Care Register for Social Welfare)
Finland	Research Services at Statistics Finland
Finland	Finnish Social Science Data Archive
Finland	Avohilmo, Register of Primary Health Care Visits
Finland	THL Biobank
Finland	Findata
Finland	FinSote, Health Interview Survey
Finland	Finnish Cancer Registry
Belgium	Belgian human genomics project
Belgium	Belgian Cancer Registry
Belgium	Health Interview Survey
Belgium	Health Examination Survey
Belgium	Statistics Belgium

One way to deal with the high fragmentation of health data in Germany (or Spain) could be to focus the research question in specific regions and try to gather and link individual level data at a regional level. For Germany, different regional registries relevant for the cancer use case are being identified. These registries collect standardised data sets of the cancer patients in a defined region and data could be connected to other registries (e.g. residential registry, data from social security or statistics agencies etc.) upon a specific research question.

3. Analysis requirements

3.1. Types of analysis envisaged

Cancer risk is determined by a complex interplay of genetic and environmental factors. Polygenic risk scores (PRS) method[5-7] provides a personalized genetic susceptibility profile that may be

leveraged for cancer prediction. For that, the effects of cancer-related loci identified from previous Genome-Wide Association Studies (GWAS) will be aggregated. Specifically, the Single Nucleotide Polymorphisms (SNPs) and the adjusted effect sizes will be applied to our limited set of data and the PRS will be calculated. The added predictive value of integrating modifiable lifestyle-related risk factors with cancer-specific PRS will be tested. The first step will be to assess the interaction between a PRS and a particular environmental variable with the risk of a particular cancer. A genome-wide gene-environment (GxE) interaction scan will be conducted between individuals' common variants, environmental exposure and the risk of cancer.

3.2. Analysis development challenges

Limitations of the use case study include the small sample size, the limited power for environmental exposure-response analysis and the inability to validate the predictive models. Furthermore, for any environmental exposure variable, there could be exposure misclassification leading to a potential contamination of the reference group as well as heterogeneity among the assessment. Additionally, the SNPs used to calculate the PRS in the study will be identified from previous GWAS of cancer, in which we may not know if proportions of samples have been exposed or not.

The functional requirements will be expressed in terms of the analysis elements that will be latter included as inputs in Task 7.3 in conjunction with the functional requirements of D7.2. However, the work described in this cancer use case is raised as a theoretical exercise. The data collections described in section 2 are analysed in order to capture the available metadata (data types, data schemas) and the linkage opportunities between the different datasets but there is no aim to analyse those data to increase knowledge on cancers. The possible ways to perform the analyses should be investigated with WP5.

[5] Choi, S. W.; Mak, T. S.-H.; O'Reilly, P. F. Tutorial: A Guide to Performing Polygenic Risk Score Analyses. *Nat Protoc* **2020**, *15* (9), 2759–2772. https://doi.org/10.1038/s41596-020-0353-1.

[6] Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet* **2013**, *9* (3), e1003348. https://doi.org/10.1371/journal.pgen.1003348.

[7] Palla, L.; Dudbridge, F. A Fast Method That Uses Polygenic Scores to Estimate the Variance Explained by Genome-Wide Marker Panels and the Proportion of Variants Affecting a Trait. *The American Journal of Human Genetics* **2015**, *97* (2), 250–259. https://doi.org/10.1016/j.ajhg.2015.06.005.

4. Conclusions

The limitations of this use case in terms of data access and data analysis are summarized below. However, the cancer use case should be considered as a preparatory phase for more specific data analysis and transferability. Although it will be difficult to expand and reproduce what we do in Belgium to other countries, it will certainly create a ground for possible change by creating partnerships and enhancing community capacity and participation.

4.1. Types of analysis envisaged

The data access challenges are:

- Access the data from the data collections;
- Sharing sensitive data;
- Access to a common descriptive metadata catalogue with the available data collections;
- National versus regional data collections and registries;
- Harmonization of datasets.

4.2. Detected data analysis challenges

The main data analysis challenges are:

- Analysis of cross-borders data;
- Sample Size;
- Aggregation of the effects of cancer-related loci from different studies;
- Limited power for environmental exposure-response analyses;
- Environmental exposure misclassification;
- Heterogeneity in environmental exposure assessment;
- Inability to validate the predictive models.

Annex I – List of variables from the BELHIS18 that will be used in the cancer use case

SOCIOECONOMIC STATUS
Household composition
Household identification code, called NUM_NEMA
Date-of-Birth (month, year)
Gender
Legal marital status
Nationality
Country of birth
Age at immigration
Country of birth of the mother
Country of birth of the father
Income
Income from work (as employee or self-employed) (yes/no)
Monthly income higher the 2000 Euro (yes/no)
Household budget covers expenses (great difficulty/difficulty/some difficulty/fairly easily, easily/very easily)
Health expenses
The degree of difficulty to fit the personal care contribution into the budget (easy/hardly/impossible)
Profession
Paid job (at the moment, even when temporally interrupted) (yes/no)
Current profile for who doesn't have a paid job (Unemployed/Sickness or invalidity/Studies/Retirement/ Housekeeping without benefits/family worker/other)
Ever had a paid job (yes/no)
(Last) employed as employee or self-employed (employee/self-employed)
(Last) kind of contract of employment (indefinite period/fixed period)
Fulltime or part-time (fulltime/part-time)
Main economic activity of current (last) organisation/institution (NACE01 code)
Education level
Being a daytime student (yes/no)
Current branches of study (Primary/Secondary 1st or 2nd cycle/Secondary 3th cycle/Post-secondary not-higher /Higher, bachelor/Higher, master/Academic bachelor/licentiate, engineer or master/Doctorate/other)

Highest education leaving certificate, diploma or education degree (None/Primary/Secondary 1st or 2nd cycle/Secondary 3th cycle/Post-secondary not-higher /Higher, bachelor/Higher, master/Academic bachelor/licentiate, engineer or master/Doctorate/other)
Age when ending studies (years)
Living environment
Indicators based on regrouping by HIS team of the municipality of residence of the respondents
Province of residence (categorical)
Degree of urbanization: Each selected municipality was classified as urban, semi-urban and rural municipalities based on a combination of morphological and functional characteristics listed in "België, territoriale verscheidenheid (Mérenne, Van Der Haegen, Van Hecke, 1997)1, updated with information from the Belgian National Population and Housing Census.
Livabiality
Dwelling type (Detached house/Semi-detached/Terraced house /Apartment or flat in a building with 2 dwellings/Apartment or flat in a building with three to nine dwellings/Apartment or flat in a building with ten or more dwellings/Room or furnished studio/Residential home for the elderly/Institution for the elderly (care home, nursing home/some other kind of accommodation)
Tenure status (Owner, co-owner or usufructuary/Renter from an individual private landlord or society/Renter from a social housing association or another public association/rent-free free)
Monthly rent (price category)
Number of bedrooms (indicator for house size)
Indoor environment (household)
Disability to keep home adequately warm (Never/Occasionally/Quite often/Most of the time/Don't know)
Humidity Problem (not a problem/minor problem/fairly serious problem/very serious problem/Don't know)
Mould problem (not a problem/minor problem/fairly serious problem/very serious problem/Don't know)
Ventilation habit (daily/once or more per week but not every day/once or more per month but not every week/never/other/Don't know)
Outdoor environment
Hindrance to the neighborhood
Traffic speed (Not at all a problem/Minor problem/Big problem/Very big problem/Don't know)
Traffic intensity (Not at all a problem/Minor problem/Big problem/Very big problem/Don't know)
Waste accumulation (Not at all a problem/Minor problem/Big problem/Very big problem/Don't know)
Vandalism, graffiti and other deliberate damage to property (Not at all a problem/Minor problem/Big problem/Very big problem/Don't know)

1 Mérenne B, Van Der Haegen H, Van Hecke E. (1997). België, territoriale verscheidenheid/La Belgique, Diversité territorial, Tijdschrift van het Gemeentekrediet/Bulletin du Crédit Communal n°202

Lack of access to parks or other green or recreational public places (Not at all a problem/Minor problem/Big problem/Very big problem/Don't know)
Hindrance around home
Air Quality Nuisance
Air pollution (Not at all/Slightly/Moderately/Very Extremely/Don't know)
Odour nuisance
Industrial odour (Not at all/Slightly/Moderately/Very Extremely/Don't know)
Other odour sources (Not at all/Slightly/Moderately/Very Extremely/Don't know)
Noise nuisance
Vibrations induced by aircraft/rail/road/tram traffic or industrial sources (Not at all/Slightly/Moderately/Very Extremely/Don't know)
Road traffic noise (Not at all/Slightly/Moderately/Very Extremely/Don't know)
Noise from tram, train, tube (Not at all/Slightly/Moderately/Very Extremely/Don't know)
Aircraft noise (Not at all/Slightly/Moderately/Very Extremely/Don't know)
Industrial noise (Not at all/Slightly/Moderately/Very Extremely/Don't know)
Neighbour noise (Not at all/Slightly/Moderately/Very Extremely/Don't know)
SOCIAL HEALTH
Quantity and satisfaction with social contacts
Satisfaction with social contacts (Really satisfying / Rather satisfying / Rather unsatisfying / Really unsatisfying)
Quantity of social contacts (At least once a week/At least once a month/At least 3 or 4 times a year/At least once a year/Never)
Quality of social support
The number of people you can count on (None/1-2/3-5/6 or more)
Level of concern/interest that people show in what you do (A lot /Some/Not known with certainty/Little/Not at all)
How easy or difficult it is to get neighbour aid when needed (very easy/easy/possible/difficult/very difficult)
HEALTH STATUS
Subjective health
Five-point Likert scale (Very good/Good/Fair/Bad/Very bad)
Visual analogue scale (0=worst to best=100)
Any chronic (long-standing) illness or condition (health problem) (yes/no)
Functional limitation (strongly limited/limited/not limited)
Quality of life
Walking ability (no/slight/moderate/severe/unable)
Self-care/personal care (no/slight/moderate/severe/unable)

Daily activities (no/slight/moderate/severe/unable)	
Pain/discomfort (no/slight/moderate/severe/extreme)	
Anxiety/depression (no/slight/moderate/severe/extreme)	
Chronic diseases/long-term condition/handicap	
At least one (yes/no)	
Specific name for the condition/disease	
Daily activity restriction as a result (permanent, from time to time, not or seldom)	
Bedridden as a result (permanent, from time to time, not or seldom)	
PREVENTIVE MEASURES	
Cancer screening	**Remark**
Colorectal cancer screening via Fecal Occult Blood Test (FOBT) (yes/no)	
Colonoscopy (yes/no)	
Mammography (yes/no)	
Top ten reasons for last mammography in order of priority	
Invited for population-based breast cancer screening (yes/no)	
Preventive cardiovascular medicine	
Blood pressure taken by health professional (yes/no/don't know)	
Cholesterol test performed by health professional (yes/no/don't know)	
Diabetes management	
Blood glucose test performed by health professional (yes/no/don't know)	
Vaccination	
Influenza vaccination (yes/no/never heard of the term/don't know)	
Pneumococcal Immunisation (yes/no/never heard of the term/don't know)	≥ 45 years
Human Papilloma virus vaccination (yes/no/never heard of the term/don't know)	Females, 10-44 years

SUBSTANCE DEPENDENCE
Tobacco Consumption
Primary exposure to tobacco smoke
Having ever smoked at least 100 cigarettes in lifetime (yes/no)
Age at first whole cigarette (years)
Age at start of daily smoking (years)
Ever daily smoking (yes/no)
Number of years of daily smoking (years)
Current smoking status (Ever smoker; Daily smoker; Occasional smoker; Quitter)
Daily smoker
Number of cigarettes smoked daily (rolled or manufactured) (variable)

Number of cigars/cigarillos smoked daily (variable)
Number of pipes of tobacco smoked daily (variable)
Number of water pipe episodes daily (variable)
Type of other products smoked daily (variable)
Frequency of smoking
Intensity of smoking (3 categories)
Heavy smoking (20 or more cigarettes per day)
Tobacco dependence categories
Highly dependent smokers
Lifetime attempt of quitting smoking in daily smokers (Several times; Once; Never)
Former smoker
Time elapsed since quit smoking (<month; 1-6 month(s); 6-12 months; 1-2 years; 2-10 years; >10 years)
Occasional smoker
Past 2-year trend in smoking among occasional smokers (increase; decrease; staying the same)
Quit attempts among occasional smokers (several times; once; never)
Consumption of electronic cigarettes
Having ever tried e-cigarette (yes/no)
Current use of e-cigarette (frequency distribution)
Nicotine content
Duration of e-cigarette use
Smoker before vapour
Second-hand smoke exposure
Frequency of tobacco smoke indoors exposure (never or almost never; <1h a day; 1-5h a day; >5h a day)
Place of tobacco smoke indoors exposure (At home, work, public places or transports)
Frequency of vapors of electronic cigarette indoors exposure (never or almost never; <1h a day; 1-5h a day; >5h a day)
Place of vapors of electronic cigarette indoors exposure (At home, work, public places or transports)
Alcohol Consumption
Frequency of alcohol consumption in past 12 months (every day or almost; 5-6 days a week; 3-4 days a week; 1-2 days a week; 2-3 days a month; once a month; less than once a month; not in the past 12 months; never or only a few sips/trials in lifetime)
Frequency from Monday to Thursday (4; 3; 2; 1; none)
Daily quantity of alcohol being consumed on weekdays (Monday-Thursday) (at least 16;10-15;6-9;4-5;3;2;1;0)
Frequency from Friday to Sunday (3; 2; 1; none)
Daily quantity of alcohol being consumed on weekends (Friday-Sunday) (at least 16/10-15/6-9/4-5/3/2/1/0)
Average number of alcohol drinks per day
Frequency of risky single occasion drinking (6 or more drinks on one occasion) (every day or almost; 5-6 days a week; 1-2 days a week; 2-3 days in a months; once a month; less than once a month; not in the past 12 months; never in a lifetime)
Frequency of having 6+/4+ drinks within 2 hours
Age at first alcohol use (years)
Lifetime alcohol abstainers

Alcohol quitters (have drunk alcohol, but not in the past 12 monts)
CAGE assessment for alcohol abuse
Wanting to **c**ut down on alcohol consumption (yes/no)
Feeling **a**nnoyed that people criticized one's drinking (in this case respondent's drinking) (yes/no)
Feeling **g**uilty about others criticizing drinking (yes/no)
Having a drink upon waking in the morning to get rid of a hang-over, i.e. **e**ye-opener (yes/no)
Illegal Drugs Consumption
Having ever used cannabis (yes/no)
Age at first cannabis use (years)
Cannabis use in the past 12 months (yes/no)
Frequency of cannabis use in the past 30 days (20 days or more; 10-19 days; 4-9 days; 1-3 days)
Having ever used cocaine, crack, amphetamines, ecstasy or other similar substances (yes/no)
Recent use (past 12 months) of a drug (other than cannabis)
None (yes/no)
Cocaine use in the past 12 months (yes/no)
Crack use in the past 12 months (yes/no)
Ecstasy use in the past 12 months (yes/no)
Amphetamines use in the past 12 months (yes/no)
Methamphetamines use in the past 12 months (yes/no)
Ketamines use in the past 12 months (yes/no)
GHL/GHL use in the past 12 months (yes/no)
Heroin use in the past 12 months (yes/no)
Hallucinogens use in the past 12 months (yes/no)
Opioids use in the past 12 months (yes/no)
NPS use in the past 12 months (yes/no)
Medical drugs use in the past 12 months (yes/no)
NUTRITIONAL STATUS
Height (without shoes)
Weight (if pregnant, weight before pregnancy)
Frequency of eating fruits/vegetables or salad
Eating fruits (excluding juice) daily
Number of portions of fruit, of any sort, eating each day
Eating at least 2 portions fruit daily (6 years and over)
Frequency of eating vegetables or salad (excluding juice and potatoes)
Eating vegetables or salad (excluding juice and potatoes) daily
Number of portions of vegetables or salad (excluding juice and potatoes) eating each day
Eating at least 2 portions vegetables or salad daily (6 years and over)
Frequency of drinking 100% pure juice or vegetable juice
Drinking 100% pure fruit or vegetable juice daily

Eating at least 5 portions fruits and vegetables daily (6 years and over)
Frequency of drinking sugared soft drinks (no "light")
Drinking sugared soft drinks daily
Volume of sugared soft drinks (no "light") drinking each day
Drinking at least 1 liter of soft drinks (no "light") daily
Frequency of eating sweet or salty snacks
Eating sweet or salty snacks daily
Number of glasses of water (150 ml) drinking daily
PHYSICAL ACTIVITY
Sport, Fitness, recreational activities
Number days doing sports, fitness or recreational activities in typical week
Time spent in doing sports, fitness or recreational activities in typical week (variable: hours & minutes per day / don't know)
Number days activities to strenghten muscles in typical week
Cycling
Cycling frequency, i.e. number of days cycling (min 10' at a time) in typical week (0-7)
Cycling duration (variable: hours and minutes per day / don't know)
Walking
Walking frequency, i.e. number of days walked (min 10' at a time) in typical week (0-7)
Walking duration (variable: hours and minutes per day / don't know)
Sitting
Time spend on sitting on typical day: minutes and hours
Leisure time physical activity (during last year)
Hard training and competitive sport more than once a week; Jogging and other recreational sport or gardening at least 4h per week; Jogging and other recreational sport or gardening at most 4h per week; Walking, bicycling or other light activities at least 4h per week; Walking, bicycling or other light activities at most 4h per week; Reading, watching TV or other sedentary activities; don't know
At risk due to a lack of leisure time physical activity

Annex II – Accessibility information for Belgian data collections

INDICATORS	Question	Health Interview Survey	Health Examination Survey	Belgian Cancer Registry	Genomic data registry	Statbel
Data access	Do you provide access to individual and/or aggregated data (for third party users)?	/ Individual / Aggregated	/ Individual / Aggregated	/ This doesn't apply to this data infrastructure	/This doesn't apply to this data infrastructure	/ Individual / Aggregated
	How is the data accessed (e.g. template of how to request data, access request form (link), flow chart)? Please specify or provide a URL.	https://his.wiv-isp.be/nl/SitePages/Procedure_gegevens2018.aspx	https://his.wiv-isp.be/nl/SitePages/Procedure_gegevens2018.aspx	Not applicable	No mechanisms are in place	https://statbel.fgov.be/nl/over-statbel/wat-doen-we/microdata-voor-onderzoek
	Are the conditions of access published?	yes	yes	No	No	Yes
	Is it possible to extract the data from the data infrastructure (e.g. download) or do they have to stay in the data infrastructure?	Once given access, the requested datafile is secured and transferred	Once given access, the requested datafile is secured and transferred	Certain BCR employees can extract data from the data infrastructure. No external users can access the infrastructure.	Data can currently not be extracted from the data infrastructure	No, the individual level data or aggregated data is transferred in a secure manner
	If we cannot extract the data, is there a safe space to analyse the data?			Yes, through a secure, remote environment	No	No
Registration	Do third party users have to register to the data infrastructure and have an account in order to access the data?	Yes	Yes	No	/This question doesn't apply to this data infrastructure	No
Legal approval	Does the requestor need a privacy and/or legal approval to access the data?	Yes	Yes	Yes	/ I don't know	Yes
	How long does it take to provide access to the requested data to the researcher after the query has been launched or the application for access has been submitted?	Around 6 weeks, if all goes well. Longer if the request has to go through the Information Security Council, then it is variable	Around 6 weeks, if all goes well. Longer if the request has to go through the Information Security Council, then it is variable	Very variable. Depends on the request, the need to link additional data sources etc.		3 weeks

Annex III – Accessibility information for Finnish data collections

INDICATORS	Question	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits
Data access	Do you provide access to individual and/or aggregated data (for third party users)?	/ Individual / Aggregated	/ Individual / Aggregated	Individual	/ Individual / Aggregated	/ Individual / Aggregated	/ Individual / Aggregated	/ Individual / Aggregated	/ Individual / Aggregated	Aggregated
	How is the data accessed (e.g. template of how to request data, access request form (link), flow chart)? Please specify or provide a URL.	https://thl.fi/en/web/thl-biobank/for-researchers/sample-collections/national-finhealth-study	https://thl.fi/en/web/thlfi-en/statistics-and-data/data-and-services/data-requests-and-analytical-services	https://www2.tilastokeskus.fi/tup/mikroaineistot/ohjeita_tutkijalle_en.html https://www2.tilastokeskus.fi/sivusto/lomakkeet/index_en.html	https://services.fsd.tuni.fi/index?lang=en	https://thl.fi/en/web/thl-biobank/for-researchers/application-process	Via remote access environment Kapseli	Through Findata	https://syoparekisteri.fi/palvelut/tietopyynnot/ https://findata.fi/en/	https://sampo.thl.fi/pivot/prod/fi/avopika/pikarap01/summary_kaynnitkkvko
	Are the conditions of access published?	No	Yes	Yes	Yes	Yes	Yes	This doesn't apply to this data infrastructure	Yes	Yes
	Is it possible to extract the data from the data infrastructure (e.g. download) or do they have to stay in the data infrastructure?	No	Yes	The data has to be handled over a remote access system. Researchers can download aggregated data and results from the remote access system	Customers download the data for themselves	A copy of the specific data is provided to researchers with approved research application and signed MTA	Not possilble	No	It is possible to extract from Findata	Yes
	If we cannot extract the data, is there a safe space to analyse the data?	No	Yes	Yes	This doesn't apply to this data infrastructure	This doesn't apply to this data infrastructure	Yes	No	Yes	This doesn't apply to this data infrastructure
				https://www2.tilastokeskus.fi/tup/mikroaineistot/etakaytto_en.html		not applicable	https://findata.fi/en/kapseli/		https://findata.fi/en/	https://thl.fi/fi/tilastot-ja-data/aineistot-ja-palvelut/avoi

										n-data#Peruster veydenhuolto
Registration	Do third party users have to register to the data infrastructure and have an account in order to access the data?	This doesn't apply to this data infrastructure	I don't know	Yes	Yes	This doesn't apply to this data infrastructure	Yes	This doesn't apply to this data infrastructure	No	No
Legal approval	Does the requestor need a privacy and/or legal approval to access the data?	Yes	I don't know	Yes	No	Yes	Yes	Yes	Yes	Yes
	How long does it take to provide access to the requested data to the researcher after the query has been launched or the application for access has been submitted?	6-12 months		Depending on the type of data , 1 - 6 months	For most of the cases the customer can download the requested data right away (automatic authentication and approval). If the dataset requires permission from the data depositor, it may take from a few days to a couple of weeks.	Depending on many factors, because access requires approval of application and signed MTA.	Depends on the case, current median time is 68 days	6-12 months	The permission process takes multiple months. When the requester has the legal approval, 2-4 weeks to get access to the data.	

Annex IV – Source of metadata catalogue in Finland

	FinHealth 2017 Survey	The Care Register for Social Welfare	Research Services at Statistics Finland	Finnish Social Science Data Archive	THL Biobank	Findata	FinSote	Finnish Cancer Registry	Avohilmo, Register of Primary Care Visits
Produce or collect metadata for all their data.	No		Yes we do, but some are only readily available within Statistics Finland and can be obtained only by asking separately	Yes, a description of the used format and metadata we provide: https://www.fsd.tuni.fi/en/services/depositing-data/ddi/	We produce metadata for different datasets, as well as collect documentations from research data returned to the biobank.	Data controllers expected to provide the data descriptions in Aineistoeditori (a tailor-made tool)	No	yes, https://aineistokatalogi.fi/catalog/studies/21085403-7be8-4f93-bf05-231518c642a0 https://cancerregistry.fi/services/information-requests/	
Public metadata catalogue service?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
		Sosiaalihuollon hoitoilmoitusrekisteri 1995- (Sosiaalihilmo)	https://taika.stat.fi/en/	https://services.fsd.tuni.fi/index?lang=en	https://thl.fi/en/web/thl-biobank/for-researchers/sample-collections	https://aineistokatalogi.fi/catalog	https://aineistokatalogi.fi/catalog	https://aineistokatalogi.fi/catalog/studies/21085403-7be8-4f93-bf05-231518c642a0	https://www.julkari.fi/bitstream/handle/10024/138288/URN_ISBN_978-952-343-346-5.pdf?sequence=1&isAllowed=y

Acronyms and Abbreviations

- BCR - Belgian Cancer Register
- D – deliverable
- GWAS – Genome-Wide Association Study
- HES - Health Examination Survey
- HIS - Health Interview Survey
- M – Month
- PRS - Polygenic Risk Score
- SNP – Single Nucleotide Polymorphism, i.e. differences in the nucleotide composition at single positions in the DNA sequence.
- WGS – Whole-genome sequencing
- WP – Work Package

HEALTHYCLOUD
Health Research & Innovation Cloud