**APPLICATION TO THE ESTONIAN COMMITTEE ON BIOETHICS AND HUMAN RESEARCH**

**FOR ETHICAL EVALUATION OF THE RESEARCH PROJECT**

|  |
| --- |
| **1. Name of the study (in case of an application in English, the name of the study in Estonian is required in parallel)** |
| Predicting risk of non-communicable disease (PREDICT)Mittenakkushaiguste riski ennustamine (PREDICT)The application is submitted in English, as the researchers are from abroad and do not speak Estonian. |
| **2. The main purpose of the study (up to 450 characters / 0.25 pages) (in case of an application in English, the main purpose of the study should be provided in Estonian, too)**  |
| Modern medicine has made significant advancements in improving survival for cancer patients, particularly in early detection and novel therapies. Cancer patients in Europe now live nearly six times longer post-diagnosis compared to 40 years ago. Despite these advancements, there is a noticeable lag in extending individuals' "health span"—the time they maintain a high quality of life in good health – without suffering from chronic disease. Up to 40% of cancer cases can be prevented through lifestyle modifications. However, appreciation of how specific behaviours impact individual’s risk of cancer is limited. Existing risk calculators are not able to inform individuals on the total long-term benefits of lifestyle changes across all cancer types. The overarching aim of this project is to establish a unified model for long-term assessment of non-communicable disease (NCD) risk that outputs easy-to-understand risk metrics. Amongst NCDs, we will initially consider any cancer, CVD and diabetes, but may subsequently incorporate additional disease endpoints. **The main purpose of the study is to:**1. Establish and validate disease-specific risk models based on demographic, behavioural, and anthropometric data.
2. Enrich the standard risk models with risk information from polygenic risk scores and relevant biomarkers.
3. Integrate disease-specific models to predict risk of overall NCD and disease-free survival and evaluate model validity.

**Main purpose of the study in Estonian:**Moodne meditsiin on teinud olulisi edusamme vähipatsientide elulemuse parandamisel, eriti varase tuvastamise ja uute ravivormide osas. Vähiga patsientide eluiga Euroopas on nüüd peaaegu kuus korda pikem pärast diagnoosi saamist võrreldes 40 aasta taguse ajaga. Hoolimata nendest edusammudest on märgatav mahajäämus inimeste tervena elatud aastate pikendamisel – aja, mille jooksul nad säilitavad hea tervise ja elukvaliteedi, ilma krooniliste haigusteta. Kuni 40% vähijuhtudest on võimalik ära hoida elustiili muutuste abil. Siiski on teadlikkus selle kohta, kuidas konkreetsed käitumisviisid mõjutavad inimese vähiriski, piiratud. Praegused riskikalkulaatorid ei võimalda pakkuda inimestele teavet, millist pikaajalist kasu saadakse elustiili muutustest kõigi vähiliikide puhul. Selle projekti kõikehõlmav eesmärk on luua ühtne mudel mittenakkushaiguste (MNH) riskide pikaajaliseks hindamiseks, mis võimaldab tekitada kergesti mõistetavaid riskimõõdikuid. MNH--te hulgas käsitleme esmalt vähki, südame-veresoonkonna haigusi ja diabeeti, kuid hiljem võime lisada täiendavaid haiguste lõpp-punkte. **Uuringu põhieesmärk on:**1. Luua ja valideerida haigusspetsiifilised riskimudelid, mis põhinevad demograafilistel, käitumuslikel ja antropomeetrilistel andmetel.
2. Rikastada standardseid riskimudeleid polügeeniliste riskiskooride ja asjakohaste biomarkerite riskiteabega.
3. Integreerida haigusspetsiifilised mudelid, et prognoosida MNH-te ja haigusvaba elu üldisi riske ning hinnata mudeli kehtivust.
 |
| **3. Principal investigator(s) and their contact details** |
| **Given name(s):** *Mattias* **Last name:** *Johansson***Position:** *Scientist***Institution:** *International Agency for Research on Cancer (IARC/WHO)***Phone:** *+334 72 73 84 85***e-mail:** *johanssonm@iarc.who.int***Skype:** NA |
| **4. Other researchers involved in the study (add lines as necessary)** |
| 1. **Given name(s):** *Allison*

**Last name:** *Domingues***Position:** *Postdoctoral scientist***Institution:** *International Agency for Research on Cancer (IARC/WHO)*1. **Given name(s):** *Karine*

**Last name:** *Alcala***Position:** *Data manager***Institution:** *International Agency for Research on Cancer (IARC/WHO)*1. Given name(s): *Reedik*

Last name*: Mägi*Position: *Professor in Bioinformatics*Institution*: Institute of Genomics, University of Tartu*1. Given name(s): *Urmo*

Last name*: Võsa*Position: *Research Fellow of Functional Genomics*Institution*: Institute of Genomics, University of Tartu* |
| **5. Financing of the study** |
| **Sources of funding** | IARC/WHO |
| **Total cost of the study (amount)** | Access fees for participating cohorts in this project will be covered by IARC internal budget.  |
| **Financial compensation for the study participants (yes, no, explanation and amount)** | Not applicable |
| **Insurance provided for the study participants (yes, no, name of the insurance company and the certificate of insurance (COI))** | Not applicable |
| **6. Study period (the beginning and end dates (MM/YYYY))** |
| June 2025 to May 2029 |
| **7. Information about previous or parallel evaluation of the same study project (incl in other countries)** |
| NAThe study project have received approval from the Scientific Advisory Committee of the Estonian Biobank on 21.11.2024  |
| **8. Brief overview of previous studies on the same topic (up to 900 characters / 0.5 pages)** |
| We conducted a preliminary analysis across 16 cancer sites in UK Biobank to quantify the added predictive value of integrating cancer-specific PRS with family history and modifiable risk factors for 16 cancers. We showed that incorporating PRS measurably improves prediction accuracy for most cancers, but the magnitude of this improvement varies substantially. Kachuri L et a. Nat Comm. 2020Our collaborator (N Chatterjee) developed and validated sex-specific pan-cancer risk scores (PCRSs), defined by the combination of common risk factors and PRSs, to predict the absolute risk of developing at least one of the many common cancer types. This study demonstrated the potential of increasing the risk-benefit balance of rapidly emerging non-invasive multicancer early detection (MCED) liquid biopsy tests through risk stratification. Kim et al. NPJ Prec. Oncol. 2023Chatterjee et al. further developed a method to establish integrative risk models using incompletely measured risk indicators through Heterogeneous Transfer Learning via GMM. BioRxiv 2023**;** |
| **9. Rationale for the planned study and research questions and / or hypotheses (up to 1800 characters, 1 page)** |
| Advancements in modern medicine over the past few decades, including in early detection strategies and novel therapies, have significantly improved cancer survival. Cancer patients in Europe live nearly six times longer after their cancer diagnosis today than 40 years ago,[1] and about half of cancer patients survive for 10 years or more after their diagnosis.[2] While these advancements have increased the chronological lifespan of individuals suffering from cancer, they have not necessarily improved their ‘health span’ - how long they remain healthy and free from disease-related morbidity with high quality of life. It is estimated that up to 40% of incident cancers are preventable through health promoting behaviours such as adjustments in diet and physical activity, reduced alcohol consumption, and smoking cessation.[3] However, though most people have a sense of what a “healthy” lifestyle entails, they may not fully grasp the extent to which lifestyle affects their risk of cancer (Figure 1). Risk calculators, such as the Gail Model for breast cancer, exist to help individuals estimate their cancer risk. However, these calculators focus on individual cancers, while many risk factors such as obesity, smoking and alcohol consumption causally influence the development of several types of cancer, and typically focus on short term risk estimates (5 to 10 years). Without considering the effect of lifestyle factors over multiple diseases and cancers or for a longer time horizon, the total benefit of a shift in lifestyle cannot be assessed using existing tools alone. In addition, existing risk calculators may also fail to present risk information in a way that is understandable to individuals from the public, limiting their ability to influence health behaviours.A graph showing a diagram of people and their behavior  Description automatically generated with medium confidenceFigure 1. Health trajectories following preventive actionsIncreasing health promoting behaviour (e.g., exercising) and reducing modifiable causes (e.g., high blood pressure and obesity), have the potential to decrease the incidence of cancer by up to 50%,[4] and CVDs by 70%.[5] However, such numbers have limited relevance to motivate individuals to reduce their risk of future NCDs. We can readily describe the influence of various risk factors on the short-term risk of developing specific NCDs with the use of disease-specific risk prediction models. Such models are currently used for specific clinical purposes, for instance in (late) prevention of atherosclerotic CVD (ASCVD), and to a lesser extent, risk-based secondary prevention of cancer (i.e., screening) [6,7]. However, there are several key challenges to personalised NCD prevention that will be addressed by the CRESCENT initiative, including:**(1) Current risk prediction models focus on individual diseases.** For instance, a diabetes risk model ignores that the disease shares many risk factors with CVDs and several cancers. This means that the overall benefit of avoiding risk factors, such as obesity or smoking, that causally contribute to multiple diseases is ignored. There is no comprehensive risk prediction model that quantifies the total risk of NCDs and the potential to reduce this risk [8].**(2) Current models provide accurate risk estimates over a relatively short time-period, typically over 5 to 10 years.** Such estimates are used by healthcare professionals to guide early detection-based interventions in individuals who are already at high risk for specific diseases. However, such short-term risk estimates do not inform on the long-term benefits of risk reduction through preventive actions earlier in life.**(3) Personalised prevention lacks evidence-based risk communication and contextualised guidance.** Successful personalised prevention requires carefully developed health counselling and contextualised interventions. Current strategies do not consider individual values or risk literacy, nor social determinants of health and disease, both of which are crucial to determine personal health competencies and self-efficacy.**(4) Prospective intervention studies evaluating the effectiveness of implementing risk tools for disease prevention and risk reduction are lacking.** Due partially to the lack of a comprehensive prediction model for NCD risk, very few initiatives have translated risk prediction tools into disease prevention or evaluated their efficacy in prospective interventional studies that measure disease-related endpoints.**(5) Policy makers lack an evidence-based tool that assesses population impact from different actions to allow prioritisation of prevention strategies.** This is partially owing to the lack of comprehensive tools that can estimate health consequences following implementation of different disease prevention policies.**We hypothesise** that addressing these challenges will require a comprehensive personalised prevention framework, including i) robust prediction models that can accurately estimate overall long-term NCD risk ii) using understandable risk figures to iii) support evidence-based and effective health counselling that can be iv) implemented and adapted to the individual, community, and country to prevent NCDs. Establishing such a risk prediction framework will improve our ability to quantify the impact of disease preventive actions across common NCDs and may also be useful to identify individuals who may benefit from emerging early-detection modalities, such as multi-cancer early detection (MCED). References:1. R. Berman *et al.*, “Supportive Care: An Indispensable Component of Modern Oncology,” *Clin Oncol*, vol. 32, no. 11, pp. 781–788, Nov. 2020, doi: 10.1016/J.CLON.2020.07.020.
2. H. Strongman *et al.*, “Medium and long-term risks of specific cardiovascular diseases in survivors of 20 adult cancers: a population-based cohort study using multiple linked UK electronic health records databases,” *Lancet*, vol. 394, no. 10203, p. 1041, Sep. 2019, doi: 10.1016/S0140-6736(19)31674-5.
3. I. Soerjomataram *et al.*, “Cancers related to lifestyle and environmental factors in France in 2015,” *Eur J Cancer*, vol. 105, pp. 103–113, Dec. 2018, doi: 10.1016/J.EJCA.2018.09.009.
4. The World Health Organisation. Cancer Prevention. Available at: <https://www.who.int/activities/preventing-cancer>
5. Yusuf S, Joseph P, Rangarajan S, *et al*. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. Lancet. 2020 Mar 7;395(10226):795-808. PMID: 31492503.
6. Tokgozoglu L, Torp-Pedersen C. Redefining cardiovascular risk prediction: is the crystal ball clearer now? Eur Heart J. 2021 Jul 1;42(25):2468-2471. PMID: 34120165.
7. Tammemägi MC. Application of risk prediction models to lung cancer screening: a review. J Thorac Imaging. 2015 Mar;30(2):88-100. PMID: 25692785.
8. Scarborough P, Harrington RA, Mizdrak A, Zhou LM, Doherty A. The Preventable Risk Integrated ModEl and Its Use to Estimate the Health Impact of Public Health Policy Scenarios. Scientifica (Cairo). 2014;2014:748750. doi: 10.1155/2014/748750. Epub 2014 Sep 25. PMID: 25328757.
 |
| **10. Research methodology (up to 1800 characters, 1 page)** |
| Data to be obtained: To allow time-to-event modelling of any cancer, CVD and diabetes, we request access to data on all cohort participants (no restrictions) to access demographic, behavioural, anthropometric, and clinical data. Data will be harmonised among all participants from the Estonian Biobank and other cohorts (EPIC, UKB, HUNT)Design and analysis plan:We will focus on three disease groups that are jointly responsible for more than 60% of the total disease burden in Europe – CVD (leading cause of death, with 37% of all deaths in 2017), cancer (26% of all deaths), and type-2 diabetes (2%). This focus is justified because these NCDs share many modifiable causes that can be intervened upon and measured. Developing the framework for building the NCD prediction model will allow for incorporation of additional NCDs in the future, such as dementia. Furthermore, cancer is heterogenous and the extent to which it can be prevented depends on the site of origin. For instance, lung cancer is highly preventable whereas prostate cancer is not. We will still incorporate prostate cancer and other ‘less preventable’ cancers in the NCD risk model as they represent a significant portion of the cancer burden and may benefit from risk-informed secondary prevention (i.e., early detection). The modelling process will involve *i)* fitting prediction models for specific diseases (individual cancer types, individual CVDs, and diabetes), *ii)* enriching the models by incorporating biomarkers information (genetics, proteomics, blood biomarkers, etc.), and *iii)* integrating these models to predict risk of overall NCD, and disease-free survival. i) For many NCDs there are already well-validated models that can be used to predict disease-specific risk. However, the methods for establishing these models vary between diseases, and whereas that does not invalidate the models, it will be important to ensure that each disease is modelled using a common Cox-based framework to allow estimating the total disease risk. We will therefore refit/recalibrate each disease model using Cox-regression as needed. The diseases include the following ICD codes: all malignant neoplasms excluding non-melanoma skin cancer for prevalent and incident cases (ICD-10 C00-C97 excluding C44); Type 2 diabetes mellitus for prevalent and incident cases (ICD-10: E11); Acute myocardial infarction, subsequent myocardial infarction, complications following acute myocardial infarction, other acute ischaemic heart diseases, chronic ischaemic heart disease, cerebral infarction, stroke not specified as haemorrhage or infarction (I21, I22, I23, I24, I25, I63, I64).ii) We will incorporate additional risk indicators using a data integration and transfer learning methodology developed by our collaborator Nilanjan Chatterjee that allows building models in the setting of partially observed data. The statistical framework entitled Heterogeneous Transfer Learning via GMM has recently been used to establish risk models for multiple diseases in UK Biobank using the Olink proteomics data available on 50,000 research participants, whilst making use of the full cohort database of 500,000 individuals with standard risk indicators. We may also consider incorporating single important, but unmeasured, risk indicators using the Individualised Coherent Absolute Risk Estimator (iCARE) method (Chatterjee), which predicts risk by integrating multiple sources of data (risk factor associations with disease, population risk factor distributions, and population rates of disease and mortality). iii) Finally, we will compute cumulative incident functions of each NCD and overall NCD risk by combining the cause-specific hazards estimated in the disease-specific models. Initially, this will involve using simple disease-specific models with key risk indicators (e.g. age, sex, BMI, tobacco and alcohol use) to ensure that the overall risk estimates are well calibrated in the presence of strong competing risks before introducing more complex models. All risk models will be externally validated with respect to their calibration and discrimination. |
| **11. Study sample and description of recruitment method. Information and consent forms, questionnaires and tests should be submitted as annexes to the application.** |
| **Sample size, inclusion of control groups**  | Total number of participants without exclusion in the Estonian Biobank (around 212,000), age >18 years old. |
| **Who recruits and how/where/by whom is informed consent obtained? (if applicable)** | No re-contact or interaction with participants is required. Secondary analysis on available data will be performed. |
| **How and from whom are the subjects selected (sampling frame)? What are inclusion or exclusion criteria of subjects?** | No re-contact or interaction with participants is required. Secondary analysis on available data will be performed. |
| **Type of interventions (physical, mental or data, including special categories of personal data)** | NA |
| **Burden on the subject (methods of contact, number of visits, type and number of procedures, repetition of invitations, etc.)** | NA |
| **12. Issuing of tissue samples to third parties (RNA, DNA, plasma etc)** |
| **The number of gene donors whose tissue samples will be issued and the types of tissue samples to be issued**  | NA |
| **The amount of tissue sample to be issued per one gene donor** | NA |
| **The entity to whom tissue samples will be issued (country, institution, address)?** | NA |
| **What will be done with the residue samples (will the residue samples be destroyed or sent back to Gene Bank)?** | NA |
| **13. Analysis of the ethical aspects of the study (3600 characters, up to 2 pages).****All research involving human subjects must be carried out in compliance with ethical requirements, in particular the principles of respect for autonomy, charity and the prevention of harm, and justice.** (<https://www.coe.int/en/web/bioethics/guide-for-research-ethics-committees-members>). The researchers involved in data analysis are employed by the International Agency for Research on Cancer (IARC/WHO). IARC/WHO is committed to upholding the highest ethical standards in all research activities, including those involving human biological samples, data protection, and privacy concerns. While IARC/WHO is not bound by national or regional data protection laws due to its international status, it ensures that human data is handled in line with internationally recognised data protection frameworks. This includes compliance with the Personal Data Protection and Privacy Principles for UN System Organisations (UN-HCLM 2018) and the [IARC Data Protection Policy](https://www.iarc.who.int/wp-content/uploads/2024/07/IARC-Data-Protection-Policy.pdf). In this study, researchers will access data remotely from the Estonian Biobank. The following ethical principles have been specifically considered in the planning of this study:1. **Respect for Individual Autonomy**

Respect for autonomy is a core ethical tenet of this study. All participants in the Estonian Biobank have provided informed consent to participate in research, including the use of their health data and genetic information. The data will be accessed in pseudonymised form, ensuring participant identities are protected. No re-identification or re-encoding of the data will be permitted. Researchers will access the data remotely, and it will be used exclusively for modelling and predicting risk of cancer and other non-communicable diseases (NCDs), with an emphasis on the long-term effects of lifestyle changes.1. **Non-Maleficence (Do No Harm)**The principle of non-maleficence underpins the entire project design. The study aims to develop and validate risk models for cancer, cardiovascular disease (CVD), and diabetes using demographic, behavioural, and anthropometric data. Advanced methodologies, including polygenic risk scores and biomarker integration, will be employed to ensure scientific rigour and accuracy. The study is designed to minimise risk and maximise benefit by providing insights that can inform personalised prevention strategies. No interventions will be carried out directly on participants, and all data will remain securely managed, significantly reducing the potential for harm.
2. **Justice**

The Estonian Biobank represents a population-based cohort encompassing a diverse cross-section of Estonian society, including variation in age, sex, and socioeconomic background. This diversity enhances the generalisability of the findings. By developing integrated risk models that estimate the likelihood of various NCDs and predict disease-free survival, the research aims to support public health by improving risk communication and prevention strategies across the population. Ultimately, the study seeks to deliver equitable health benefits, contributing to the reduction of NCD burden in a fair and inclusive manner. |
| **13 a Human subjects**  |
| **Assistance questions** | **No** | **Yes** |
| **Are people the object of research?** |  | Secondary analysis of pseudonymised individual level data will be performed. All of the participants have joined the Estonian Biobank voluntarily and given informed consent for data analysis for the purpose of scientific research. Nobody will be discriminated against upon joining or for the fact of being a participant. The participant can withdraw consent at any time. Gene donors can prohibit the use of other databases containing records of the donor.  |
| **Are the study participants vulnerable individuals or groups?** | **X** |  |
| **Does the study include persons who cannot themselves give informed consent to participate in the research (incl. persons with limited active legal capacity)?** | **X** |  |
| **Are the study participants children/minors?** | **X** |  |
| **Are the study participants patients?** | **X** |  |
| **Does the research involve collection of biological samples? Are human biological samples intended for export to a third country (**[**https://www.aki.ee/et/teenused-poordumisvormid/andmete-edastamine-valisriiki**](https://www.aki.ee/et/teenused-poordumisvormid/andmete-edastamine-valisriiki)**)** **(**[**https://www.aki.ee/en/guidelines/transfer-personal-data-foreign-country-0**](https://www.aki.ee/en/guidelines/transfer-personal-data-foreign-country-0)**) or import them from another country to Estonia?** | **X** |  |
| **13 b Personal data and datasets** |
|  | **No** | **Yes** |
| **Are personal data collected or analyzed in the study, including special categories of personal data?** |  | Pseudonymised previously collected individual level data on study participants will be analyzed, including risk factor information, genomic data, and disease information.  |
| **Does the research involve systematic monitoring of an individual, the collection of his or her data profile, or a large-scale processing of data of special categories and /or sensitive data, or the use of (intrusive) data processing techniques in a covert way (eg survival surveys, monitoring, surveillance, audio and video recording, geolocation, etc.) or any data processing that may harm the rights and freedoms of the data subject?** | **X** |  |
| **Is there a plan to analyze previously collected personal data?** |  | Pseudonymised previously collected individual level data on study participants will be analyzed, including risk factor information, genomic data, and disease information.  |
| **Is there a plan to analyze publicly available data?** | **X** |  |
| **Is there an intention to transfer personal data or provide access to personal data to third countries****(**[**https://www.aki.ee/et/teenused-poordumisvormid/andmete-edastamine-valisriiki**](https://www.aki.ee/et/teenused-poordumisvormid/andmete-edastamine-valisriiki)**)?****(https://www.aki.ee/en/guidelines/transfer-personal-data-foreign-country-0)** | X |  |
| **Will personal data be destroyed / anonymised at the end of the research?** |  X |  |
| **13 c Other ethical issues** |
| **Can conducting research involve ethical risks not described above?** | **X** |  |
| **14. Complete in case the research is based on data from a database and/or register** |
| **Name of database and/or register** |  The Estonian Biobank database |
| **Purpose of the processing of personal data** | The purpose of processing is to develop and validate risk models for non-chronical diseases.  |
| **List of variables and period for which data are collected (in annex if necessary)****We will analyze data in the following categories*** **Genealogical data (family history of medical conditions spanning four generations, if available) including:**
* Cancers (breast, brain, bowel, gastrointestinal, liver, lung, lymphoma, myeloma, prostate, and stomach/gastric)
* Cardiovascular disease
* Diabetes
* High blood pressure
* High cholesterol
* **Educational and occupational history:**
* Highest degree completed
* Occupational history
* Occupational exposures (paint, asbestos, pesticides)
* **Lifestyle data including:**
* Physical activity
* Smoking information (status, intensity, duration, quit years in former smokers)
* Alcohol information (status, intensity, duration)
* Dietary information (processed meat consumption, fruit and vegetable intake, salt intake)
* Other dietary information (FFQ etc)
* residence data at county level (to assess indirectly pollution exposure)
* **Women’s health information including:**
* Age at menarche
* Menopausal status
* Use of menopausal hormonal therapy
* Contraception use (type and duration)
* Parity
* Age at first live birth
* **Other health measurements:**
* Weight, height
* HbA1c
* PSA
* Blood pressure
* Cholesterol levels
* Pulse rate
* Triglycerides
* Glucose
* Uric acid
* C-reactive protein
* **Genetic data**
* **Clinical information on incident and prevalent disease including:**
* Allergic conditions
* Diabetes
* High cholesterol
* Hypertension
* Hepatitis C/B
* Non-viral liver disease
* GERD
* Chronic kidney disease
* COPD
* Cardiovascular disease
* Cancers
 |
| **15.** **Description of personal data protection measures, including data storage, security and erasure, including date of erasure of data and / or code key (up to 1800 characters, 1 page).** |
| **Describe and justify the storage of data collected for the study and the deadline for storage.** | All data used in this study will be accessed remotely through the SAPU environment in a pseudonymised format. No identifiable personal data will be transferred or downloaded. Researchers from IARC/WHO will access the data through a secure, encrypted platform provided by the University of Tartu, which includes strict access controls and user authentication.IARC/WHO is committed to safeguarding personal data in accordance with international standards, including the UN Personal Data Protection and Privacy Principles (2018) and the IARC Data Protection Policy. Although not subject to national legislation, IARC applies rigorous internal protocols to ensure data privacy, confidentiality, and security.Data will be used solely for the purpose of this study and will not be depseudonymised. No copies of the data will be stored outside the secure platform. Upon completion of the study, access rights will be revoked in line with the Estonian Biobank’s policies. The code key linking pseudonyms to personal identities is held exclusively by the Estonian Biobank and is not accessible to IARC researchers at any stage.At the end of the project, a data audit is conducted to decide which data to keep or archive and which to delete. After that, the SAPU environment is deleted.Deadline: May 2029 |
| **Describe the process and means of pseudonymisation of personal data.** | Pseudonymisation is performed by Estonian Biobank. |
| **Is there a plan to de-pseudonymise gene donors’ data?** | 1. **Please specify the number of gene donors whose data will be de-pseudonymised.**
2. **Please explain the reason for de-pseudonymisation.**
 |
| **Is there a plan to transport personal data? Please describe how data protection is ensured.** | Personal data will be analysed in SAPU. |
| **Describe how the data are protected against unauthorized or unlawful processing.** | The Sensitive Data Analysis Platform or SAPU is an environment provided by the University of Tartu High-Performance Computing Centre, where analysts and programmers can work on sensitive data <https://docs.hpc.ut.ee/public/services/SAPU/>The environment reduces the risk of possible unauthorized copy, transfer, or retrieval of sensitive data from the machines, providing a higher class of security than that of a standard high-performance cluster.SAPU is an isolated environment where:•The machine has no direct access to the internet.•Complete network isolation based on firewall rules.•Access to the machine is possible only through a virtual desktop environment. •Analysts can move files using object storage, which saves anything moved.•Moving files out requires approval from the data owners' side.•The monitoring layer and the server record all actions taken.•SAPU supports most research tools available to scientist. |
| **I confirm that all researchers are aware of the ethical and personal data protection requirements of the project.** |
| **Signature of the principal investigator** | **Date of application**19.05.2025 |
|   |  |
| **EBIN ID of the application****(fills by the assessor)** |
|  |

**List of additional documents:**

**1. CV of the principal investigator**