|  |
| --- |
| **PhD Fellowship Strategic Basic Research  PROJECT OUTLINE (max. 12 A4 pages)** |

**Review primer:**

1. Context story – the flow of the context is a work in progress but what are your thoughts on the **current conceptual flow**? Do not worry about specific comments (like grammar) because there is a lot of specific work that needs to be done (unless you have a relevant specific comment like a citation suggestion or something similar).
2. Objectives – These objectives are a good reflection of how I currently want to present the project. Their descriptions need some work but I am curious if there is feedback about **if these aims are good (or at least compelling in the eyes of a potential jury**)?
3. Methodology – Again, not fully formed but the **cornerstone ideas are there**. I would love suggestions about any specific content that you think I should focus on more in-depth or important concepts or information that I am missing.
4. Economic added value – This is the most rough of the sections. I just threw a bunch of stuff from my VITO proposal and presentation here. I have put 0 effort into blending this yet. For comments here, please feel free to **suggest potential products, companies, or other applications** that I can discuss in this section. I think I will not struggle with this too much but name dropping **specific Flemish** stuff is a really good idea and it is an area that I am not terribly well informed yet…

Also for context:

* RED colored text = places that I need to go back and fix/add later [i.e. it is a section that is incomplete or I need to revisit because I got stuck or don’t like it] (so RED are artifacts from my writing approach, that’s all really).

**1. Rationale and positioning with regard to the state-of-the-art**

Research Context

As our understanding of genomics deepens, the role of Personal Genome Sequencing (*PGS*) in healthcare becomes increasingly vital. At the time of writing, there are multiple domains of clinical practice where patient personal genomic sequence (*PGS*) data is used to inform medical decision making1,2. For over a decade, the scaling of approaches that utilize PGS data to generalized clinical practice has been right around the corner3. Unfortunately, barriers to expanding PGS usage in clinical practice remain4. The proposed PENGQUIN project aims to address one group of these challenges -- those presented by the digital PGS data itself. We first identify the key concerns presented by the three constituents of a clinical ecosystem, those of the **patient**, **clinician**, and **healthcare administration**. From the perspective of **patients**, there are **privacy and transparency concerns about their PGS data**. From the perspective of **clinicians**, the **accessibility and interoperability of the PGS data for various tests and assessments** is important. From the perspective of the **healthcare administration**, the **cost of generating and storing digital patient PGS data as well as the associated privacy consideration presented by storing and accessing such data** carries high importance. With these crucial considerations as a foundation, we aim to develop a citizen-centric proof-of-concept framework that through the leveraging of the Solid protocol and development of PGS querying and data interface strategies, addresses these unique challenges and drives the development of implementable products and/or companies that expand clinical PGS data usage in Flanders, Belgium, the European Union, and beyond.

For our proposed project to have an immediate impact on medical practice, PGS data use cases must both exist and exhibit improved outcomes compared to current practices that do not utilize PGS data. At present, there are three major medical fields that preliminarily show readiness for and benefit from greater use of PGS data, these being drug development and prescription, cancer diagnosis and treatment, and rare genetic disease identification and treatment.

Genomic variation has been shown to impact many aspects of drug prescription and treatment efficacy5. More recently, genomics has also become informative for new drug development6. In relation to clinical practice, many studies have documented genotype correlation with unintended drug responses for commonly prescribed drugs such as warfarin, ..., ..., etc. Clinically testing for patient genetic predisposition to non-responsiveness or heightened risk of side effects could encourage optimal treatment planning and dosing leading to improved clinical outcomes. Furthermore, PGS has the potential to help predict, better understand, and inform optimal drug treatment plans for patients with various diseases, and represents significant improvement over the one-size-fits-all, trial and error approach popularly practiced​7.

For cancer patients, the use of PGS data has been shown to similarly augment existing strategies of diagnosis and treatment. Studies have established that PGS data analysis can improve tumor diagnosis capabilities, development of targeted, personalized therapies, ....

and allow for the development of more effective treatments8​​.

Rare genetic diseases are one of the most obvious areas of medicine that PGS data has proved particularly useful in improving disease identification and treatment. Because an estimated 80% of rare diseases are of genetic origin9, and patients searching for diagnosis and proper treatment commonly undergo years of testing more than one institution10, new approaches for using and sharing PGS data are merited. MORE...

However, it is crucial to note that PGS is not devoid of challenges, including the potential overinterpretation of results due to a limited understanding of contextual information and the risk of genetic discrimination3.

The human genome’s growing role in healthcare, particularly at the level of the distinct, individual patient, as evidenced above, necessitates the collection, storage, and analysis of large, highly personal, PGS datasets. Currently, patient PGS data storage is achieved through an institution-centric approach. This strategy is characterized by the hospital or hospital system isolating all its stored data into one or more centralized databases that are governed and maintained solely by the owner institution. Here, centralized storage is defined as the structural organization of a data store, establishing that a single node, the system that hosts the physical data storage, has control over managing database content, access, and permissions. With this system, healthcare institutions prioritize patient data privacy over all other digital data storage conditions. This method of data storage *prioritizes patient data security and access only for authorized users within the institution’s internal network*. Crucially, this institution-centric data storage architecture, by design, allows for little to no connectivity or accessibility for users or data usage requests that originate outside of the institution’s network, largely due to the organizational design of such centralized data storage systems. Due to this isolation of patient PGS data to a single institution, data duplication leading to increased storage and data generation costs, severely limited patient data usage transparency, and severely limited potential for data sharing are observed. Collectively, these current approaches to PGS data storage inhibit the scalability of PGS data usage in the above discussed clinical applications.

To develop an improved system for PGS data storage that addresses issues presented by the current system, such challenges must first be identified. For this identification, three important perspectives are considered: the patient, the clinician, and the healthcare administration.

From all three perspectives, **privacy** is of the utmost importance. PGS data, like other personal data, is by nature private and there is open debate regarding the degree to which patients should be able to control their own personal medical data11,12. Legally, within the European Union, a core tenant of current personal data usage is *transparency*13. Protections for citizens relating to sensitive personal data is codified in the European Union’s General Data Protection Regulation (GDPR) law since 201814. PGS data, along with other health-related data, is categorized as “sensitive” data in the GDPR, carrying with it the strictest privacy safeguards. Notably, privacy is prioritized in the currently used systems, but improvements can still be made as centralized databases are cited to be prone to data leaks and ransomware attacks[c].

From the perspective of the patient, awareness and control over their personal **data usage** is an additional consideration that the current system does not do much to address. As established by the GDPR, technical infrastructure designed for the storage and accessing of PGS data requires special attention to methodologies that address patient control and/or transparency over usage of their data. Because of the architecture of centralized databases, granular permissions for specific subsets of sensitive data are challenging to implement yet could allow for a patient to view their own PGS data, thus improving their connection to their own personal data. Additionally, policy implementations such as patient ability to view and/or approve or deny consent to PGS data usage for specific instances, rather than a blanket consent at time of collection, will improve the patient’s involvement in how their data is used. (conclusion sentence?)

From the perspective of the clinician, being able to **access patient data**, **share patient data** with providers outside one medical institutions system, and not need to worry about **data interoperability** with biomedical tools, applications, or user interfaces is crucial.

**Need to merge the following 3 paragraphs with above...**

The average human genome is slightly over 3 billion base pairs in length (3 Gbp).15 During a whole genome sequencing workflow, various sequence formats, that offer different sets of information, such as FASTQ16, SAM/BAM/CRAM17,18, VCF19, and others are produced. With respect to clinical practice, the most useful of these files is the VCF file[c]. VCF files are typically between 100-1000s MB (0.1-1 GB) in size within computer memory when compressed and represent ~106 nucleotide positions of an individual’s PGS. Despite being orders of magnitude smaller than other PGS representations, such datasets still present storage and querying problems due to their size. Additionally, the complexity introduced by storing such data in ways conducive to protecting personal privacy further compound the challenge presented. (something here about Resource Description Framework (*RDF*) and how it could be useful...)

Despite such prerequisites, there is a conspicuous gap in the current scientific discourse around the development and implementation of citizen-centric, consent-driven personal genome data sharing.

(3) **Querying**. While specific use-cases will be discussed later in this proposal, within a datastore, (want to say: only some data is relevant within the whole store... so we only want those bits). To identify the data that is relevant to a particular use case, a querying mechanism is required. The process of querying data is further complicated when data access permissions vary for different subsets of data and data is being searched for across different datastore locations.

From the perspective of healthcare administration, the cost of generating, storing, and maintaining patient PGS data, with the necessary privacy measures and monitoring, is monetarily demanding. ..... Cost/providing the infrastructure

Definitely more here about what is available right now ...

Clinical storage strategies

Research storage and sharing

Private companies + lack of privacy

\*\* Nothing available that prioritizes needs of physicians + patients

This gap underscores the necessity of the proposed doctoral research project. Our aim is to develop Personal Genome Pods using the Solid protocol20, a set of technical specifications that enables decentralized data storage and infrastructure for data privacy preservation alongside increased sharing capabilities. Additionally, the Solid specifications also make it possible for granulized policy application to stored data, offering specified access controls to specific data within a pod as well as the fundamental components for data usage consent requests for stored PGS data. These possibilities align well with the increasing public demand for secure, consent-driven sharing of sensitive personal data. We envisage that this project will significantly contribute to the field of personalized medicine and clinical genomic research by empowering individuals to utilize their personal genomes in healthcare and clinical practice in a manner that reinforces their autonomy and privacy.

While we continue to investigate the latest advancements and discussions in the field, we will particularly focus on the development of a data storage and sharing framework built withing the Solid ecosystem that allows for connection to existing genomics and bioinformatics tools and pipeline. We also hope our novel approach to consent-driven sharing of personal genome data contributes more broadly to ethical, legal, and public discourse regarding health data privacy and sharing protocols in general. We anticipate that our research will illuminate a path towards implementing a citizen-centric, consent-driven personal genome sharing framework in practice.

Scientific environment and strategic positioning

This project is an institutional collaboration between VITO Health, and UGhent creating a strong interdisciplinary team. The Ph.D. candidate will be supported by his supervisors with strong track records in the field of interest.

The data analysis parts of the Ph.D. candidate will be co-supervised by Dr. Ruben Taelman and Dr. Gökhan Ertaylan. Dr. Ruben Taelman and Dr. Gökhan Ertaylan will be the co-promotors of the Ph.D. candidate with additional oversight provided by Dr. Ruben Verborgh. Dr. Ruben Verborgh is a professor of Semantic Web technology at Ghent University – IMEC and will provide high level feedback regarding the aims and structure of the project. Dr. Ruben Taelman currently has a faculty position at the IDLab – IMEC at UGhent, renowned for their expertise in education, and research in decentralization, query processing, and Semantic Web technologies.

Dr. Gökhan Ertaylan is the Research Lead at Precision Health Group within VITO and the project lead in the system-level data analysis within the “I am Frontier '' study. He will supervise, support, and coordinate the study. He further developed genomics applications including polygenic risk scores for disease risk prediction and pharmacogenomics passports application for prescription of drugs in a personalized manner taking into account personal biological sensitivities and potential side effects.

The proposed project will allow the Ph.D. candidate to develop expertise as a data scientist and bioinformatician with a proficiency in areas of decentralization, query processing and software development. The Ph.D. candidate will expand his network at a national and international level by attending and presenting in research meetings, workshops, and conferences. He will register in the Data Science Hub of VITO (lead by Dr. Bart Beulens) where he will actively participate and learn from other related projects. Through FLAMES and the doctoral school of UGhent, he will develop a further expertise in data sciences and bioinformatics by following trainings and workshops for broader skill development.

VITO Health has the technological capabilities and strategic collaborations with industrial and academic partners to play a crucial role in the development of Personal Genome Sharing. However, this key role also requires extended research into the collection, storage, and processing of large genomics data sets as it pertains to privacy and ethics.

This will be the focus of this project co-funded by VITO and Ghent University. VITO health aims to investigate the optimal method of sharing personal genomes in practice across different user scenarios and technologies together with other personal health data.

The project will further benefit from the “I am Frontier” study. The “I am Frontier” study is a small cohort of 30 healthy participants, 15 females and 15 males, who were followed up over 12 months. The 30 participants were asked to donate samples such as blood, urine, and hair monthly and collect information about health parameters and lifestyle through questionnaires and wearables. Hence, “I am Frontier” presents a unique opportunity to observe and address ethical questions regarding sensitive personal information and technical challenges associated with privacy and security on a small cohort before moving to a larger population level implementation. Clear technological, ethical and technical guidelines will be explored and set not only for VITO Health, but also on a national and international level. These guidelines and workflows can be used for future cohorts by VITO Health and other institutions.

Lastly, there are two synergistic projects running in parallel with PENGQUIN at VITO Health. First one is the “Beyond the Genome: Ethical Aspects of Large Cohort Studies,” which has 3 Ph.D. candidates in total. The candidate will work synergistically with one of the PhD candidates with a background in bioethics. This ensures that both the technical and ethical aspects of genome data sharing are firmly covered while working together. Secondly, the real-life application of Pharmacogenomics Passports is being developed as part of a collaborative project with JESSA and UHasselt. This will ensure that the genome sharing functionality proposed to be developed in this project will actively collaborate and keep in mind the real-life applications during the project. Lastly, Data Science Hub is focusing on developing querying for other sorts of Health Data under the supervision of Dr. Bart Beulens which is going to run synergistically with the proposed PhD Project. Furthermore, this project is aligned with the SolidLab project at UGhent, which is investigating various research aspects of the Solid ecosystem but is not tied to a specific application domain. Concretely, the Ph.D. candidate will work closely together with 3 other Ph.D. students working in the SolidLab projects on storage and query processing challenges.

**2. Scientific research objective(s)**

Objectives

The overarching goal is to create a system that protects the privacy of personal genome data while allowing for highly performing and widely applicable genome data storage and querying. To achieve this goal, the Ph.D. candidate will investigate the use of Solid data pods to store personal genome data in a decentralized and privacy-oriented way while addressing technical challenges that come with it. The following objectives will determine the success of the PENGQUIN project.

(1) *Investigate the existing landscape of PGS storage approaches along with the storage principles and infrastructure of the Solid ecosystem to evaluate how Solid pod storage of genomic data can be achieved in a scalable, efficient manner.* [Findings will be detailed within a review paper.]

(2) *Formulate a framework for the storage and accessing of personal genomic data using Solid data pods that addresses the privacy concerns presented by PGS data for future application in clinical practice.* [This aim involves the establishment of a test dataset that will include both a single PGS as well as a set of PGSs to simulate real-world situations in which PGS storage and querying is used. Genomic data will be ingested into a personal Solid pod as Linked Data via RDF serialization using a publicly available genomic data ontology such as HL7 FHIR22 or SPHN RDF23].

(3) *Design and implement an intuitive interface for PGS data formatting.* [PGS data interoperability for human users, web applications, bioinformatics pipelines, and medical softwares is of paramount importance for practical translation of such a framework to clinical practice. Thus, starting with specific use cases, and extrapolating to more generalizable compatibility, such interfaces for data flow compatibility will be important to project success.]

(4) *Investigate query processing techniques for handling the large scale of personal genomic data stored within and across Solid data vaults. Specifically, by researching efficient Linked Data querying algorithms, and strategies for improving the performance and run-time of Linked Data traversal querying algorithms for efficient querying of PGS data for clinical and personal use.* [We should discuss this aim further in my process coaching meeting in a couple days].

**3. Research methodology and work plan**

The workflow will proceed in a relatively linear fashion. The first tasks will be the assessment of if Solid technology can accommodate efficient and effective genome storage. Simultaneously, benchmark VCF genome data will be produced to be representative of real-world VCF data for experimentation in later steps of framework development. The project will then require the definition of a genomic ontology language that will be published and made publicly available. Once a VCF ontology is established, then VCF to resource description framework (RDF) format conversion and Solid pod data ingestion will begin being experimented with. During data ingestion established data formatting standards will be integrated to allow for greater data availability for software development and scalability of this framework. Following data ingestion, query algorithms will be tested, adjusted, and created. Iterative development is likely for data ingestion methods and querying approaches due to size and complexity of VCF genomic data. Querying is planned to be possible over a single Solid pod (one VCF file) as well as over multiple Solid pods (>1 VCF files), thus iterative querying methods will also be developed.

**[remake project overview figure]**

***Figure 1 PENGQUIN project workflow.*** *White circles represent locations within the workflow where objectives will be achieved. For each of the steps in the workflow, the work packages affecting this step are also shown in parentheses (if applicable). This workflow is slightly oversimplified for ease of understanding the project.*

## Work Package 1: Review of Genome Data in Research and Healthcare practice (generation, storage, sharing, utilization) (currently underway)

The Ph.D. candidate will conduct a literature review of: the Solid decentralized data storage framework, secure/confidential data storage in decentralized environments, existing human genome single nucleotide polymorphism summary file (.vcf files) parsing approaches, methods of ingesting non-rdf data into rdf data for storage in Solid pods, the establishment of an ontology that can handle such ingestion of genomic data, rdf and non-rdf data fragmentation and fragment organization strategies, link-traversal-based query processing paradigm, and associated linked-data querying algorithms. Along with these subjects, the candidate will also investigate strategies pertaining to improving performance of query algorithms using pre-computed summaries and/or linked data short-cut summaries. This will be coupled with the representation of genomic data as Linked Data, which will either be materialized beforehand or will be mapped on-the-fly at query-time via a virtualization approach.

While performing this literature study, the Ph.D. candidate will compose a review paper on the field of storage and query processing within decentralized environments. There has not yet been published a similar research paper, and giving an in-depth overview of the current state-of-the-art techniques will help the research community work towards making personal genomics data sharing more standardized for clinical purposes.

## Work Package 2: Storing and publishing personal genomic data in a decentralized environment

In this work package, the Ph.D. candidate will set up the foundations for experimentation in later work packages. The test dataset will be constructed using publicly available Illumina platinum genome files (which files in specific?). These files will be used to provide benchmark PGS data files for testing the storage and querying approaches within a single Solid data vault and across multiple Solid data vaults. Also, this benchmark data will offer a set of realistic query workload data that mimic real-world genomic data and offer useful feedback during framework development.

We are first going to create locally host Solid pods that are representative of individual patient Solid pods using the Community Solid Server (CSS) specifications. We will then upload PGS data files from the test dataset into these patient data pods. After these simple file uploads, investigation into how PGS data can be translated into RDF format24 according to the Linked Data principles25 will be undertaken.

The goal of this work package is to investigate such a mapping process while comparing the impact of materialization and virtualization in terms of ingestion speed, storage size, and query performance. Furthermore, since data inside Solid data pods can be represented in a variety of different views, such as documents, the Ph.D. candidate will investigate different fragmentation strategies for publishing genomic data represented in RDF. These fragmentation strategies will be compared in terms of storage size, query performance, and privacy-compliance.

## Work Package 3: Interface for PGS data formatting conversions

PGS data interoperability for various applications, tools, and user interfaces is important if such a framework is to be scalable. One route to achieving this interoperability along with other potential advantages is to convert PGS data into RDF. For this, there is a need for a bidirectional mapping process, that allows PGS data to be converted into RDF, and the other way around. This mapping process could either happen as a pre-processing step during the ingestion of data into Solid data pods (materialization), or during the retrieval process if genomic data will be stored in their raw non-RDF representation in Solid data pods (virtualization).

## Work Package 4: Querying over a decentralized single individual’s genomic data then a larger population of multiple individuals’ genomic data

This work package will build upon the storage of personal genomic data in Solid data pods from the previous work package. The goal of this work package is to enable queries to be executed across PGS data that are contained within a single data pod or are spread over multiple data pods. These data will be represented as a collection of nodes that are interlinked directly or indirectly in a knowledge graph contained within the user’s personal Solid data pod. A major challenge presented by this approach is that there is a need for efficient query processing algorithms that can handle large scale decentralization. Concretely, we will build upon the link traversal query processing (LTQP) paradigm, which has been shown to be an effective method for querying within a decentralized environment such as Solid26. However, LTQP is known to currently perform sub-optimally for larger dataset sizes and complex queries, which are properties present in the PGS data space. As such, there is a need for new LTQP algorithms that can achieve the required levels of performance in terms of query execution time. Concretely, the Ph.D. candidate will assess and improve LTQP algorithms that will exploit intrinsic properties of genomic data that can help speed up query performance. Furthermore, the impact of pre-generated summaries21 will be investigated when introducing them to the decentralized environment. The working approach will then be applied to a population of PGS data in which new strategies of organizing computational architecture and utilizing metadata between genomes will be investigated to optimize storage size and query performance when querying more than one personal genome.

## 

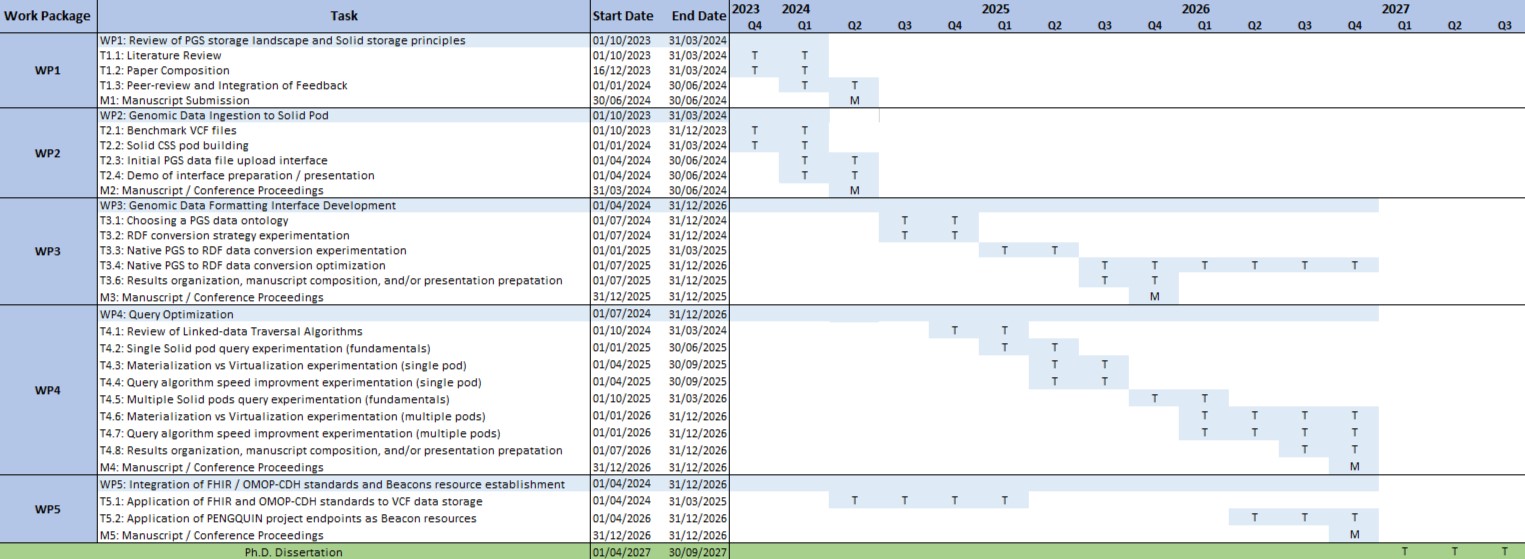
## Work Package 5: Towards practice and Embedding of the suggested Personal Genome Sharing System in the Worldwide/European/Belgian Data Sharing Landscape

The final work package be applicable to all other work packages to make the suggested work relevant in real life practice as well as allow the current project to be both improved by and applied to other applications in the future.

To avoid privacy issues and improve the communicability of sensitive medical data findings, the Observational Medical Observations Partnership (*OMOP*)27 initiative has established methods and data sharing standards that allow for the distribution of medical data findings without the fear of infringing on the privacy of sensitive medical data. Within the OMOP initiative, the genomic common data model (G-CDM)28 has been proposed as a method of improving the circulation and discoverability of genomic research results for researchers. Similarly, the HL7 FHIR Genomics Implementation initiative29 attempts to solve similar issues with genomic data sharing and discoverability without infringing on medical data privacy through the establishment of a modular data ontology. These, along with other data standard organizations such as ISO[c] or consortia such as GA4GH[c] are all viable considerations for data representation interfaces etc. We aim to integrate these data structures and discoverability standards into our PENGQUIN project architecture to encourage improved data and result sharing among the greater research and clinical community.

An overarching aim for this project is to help drive improvement and accessibility of personalized healthcare through improved PGS privacy and storage and querying efficiency. These aims align well with the international Beacon initiative30,31 to increase the availability and ease of access of genomic data for researchers globally. Thus, we aim to make the PENGQUIN Solid pod VCF data storage endpoints beacon resources that can be discoverable via the Beacon API.

Importantly, this option to contribute personal genomic data to the greater Beacon data network, thereby making it discoverable to researchers, will be controllable on an individual basis as an opt-in option. Another advantage of the Solid ecosystem in this context is the ability of any individual to change their beacon sharing preference from their personal Solid pod at any point in time. Increasing the amount of genomic data available to researchers will help contribute to the further progression of personalized healthcare efficacy for larger, and more diverse populations of people, aligning closely with our goals for this Ph.D. project.

*Planning -* The Ph.D. project consists of 5 work packages denoted as “WP”. This will be bundled into a Ph.D. dissertation for a final thesis defense. Figure 3 shows a Gantt-chart of the Ph.D. project on a quarterly basis. Each work package is split up into different tasks with a dedicated amount of time allocated to it. This will allow for a good time and project management.

***Figure 2******Gantt chart of the Ph.D. project timeline on a quarterly basis****. T denotes different tasks of a work package. M denotes different milestones, such as finishing and submitting a manuscript to a journal. Green denotes working on the thesis dissertation and defense. This is allocated as the final 6-9 months of the Ph.D. project to make all preparations for the defense.*

# Risks and mitigation (need to revisit these)

* **Dataset size:** If personal genome datasets are too large to store in a decentralized manner with acceptable query performance. If this occurs, a shift towards more centralization will be investigated in the form of summaries [3], and more expressive query interfaces will be added to Solid data vaults [7]. The range of pods over which summaries and query interfaces are defined will be configurable, which means that the trade-off between centralization and decentralization can be tuned.
* **Data availability, ethics and legislation (GDPR guidelines):** VITO Health and UGhent have an institutional agreement which regulates the legal and ethical aspects of the project. This includes confidentiality, ownership, and privacy issues. Genomics data from other sources is freely accessible online as the data is currently not legislated under GDPR guidelines.
* **Computational Load:** Various software packages will be used for brute force SAP detection. Both VITO and UGhent have access to the Flemish Supercomputing Centre for grid computing in case the Ph.D. lacks computational resources.
* **Evolving PGS Technological Advances**: Long read sequencing has been decreasing in price in recent years and offers many advantages over the currently most common form of short-read shotgun genome sequencing. With a widespread transition to long read sequencing, the way downstream genomic files, such as VCF files, are created and formatted may change. It is unlikely that such a drastic change will take place during this Ph.D. project, but if such a breakthrough occurs, the Ph.D. candidate and advisors will assess altering the input genomic files to align with the project’s goal of creating a framework useful for advancing personalized healthcare.
* **Software Dependencies:** Various software will be used for this project. Generally, software is known to exhibit issues concerning quality, versioning, and documentation. If such problems are encountered troubleshooting will include utilizing other, comparable software to accomplish the task or contacting the producer of said software for troubleshooting.

*Elaborate the different envisaged steps (experiments/activities) in your research, and motivate strategic choices in view of reaching the objectives. Describe the set-up and cohesion of the work packages including intermediate goals (milestones).*

*Show where the proposed methodology (research approach) is according to the state of the art and where it is novel. Discuss risks that might endanger reaching project objectives and the contingency plans to be put in place should this risk occur.*

*Use a table or graphic representation of the planned course of activities (timing work packages, milestones, critical path) over the 4-years grant period.*

**This needs the most work:**

**4. Strategic dimension and application potential**

**\*\*\* Disruptive Innovation --> Economic added value \*\*\***

**--Especially specifically for Flanders--**

Precision Health Group in VITO Health has the ambition to develop applications towards making personalized healthcare a reality through data-driven research. This is facilitated through having joined PhDs with academic institutes to maintain strong and relevant research. The Ph.D. candidate will actively make connections with other institutes contributing to the research field of interest. Algorithms, methodologies, and results from the Ph.D. project may be implemented in different institutions partnered with VITO, JESSA and Ghent University such as Domus Medica or Hospitals to contribute towards unveiling the personalized Genome Pods’ power in practice.

PENGQUIN, by facilitating the storing and consent driven dynamic sharing of PGS, will be poised to play a central role in developing Digital Twins of citizens/patients in healthcare and clinical practice.

Moreover, the Ph.D. project will establish a link to both the healthcare industry and computer science research by developing computational tools and methodologies which will aid in the confident storage and querying of genomic data for use in existing research aims within pharmacogenomics at VITO.

The project supports the desire to move towards a healthcare system increasingly focused on personalized genomics and preventive medicine. This is highly desired as personalized medicine offers an improved medication selection, targeted therapy, reduced adverse effects, increases patient compliance and confidence, and improves cost effectiveness [Mathur et al.]. Importantly, personalized healthcare, particularly its reliance on PGS, introduces technical and privacy challenges. Additionally, with increased use of personal genomics within healthcare, highly performing systems must be developed to store and access such data in an efficient, secure, and reliable way. Such a system presents a challenge due to the large size, complexity, and sensitive nature of personal genomic data.

This project will provide an important step toward creating such a system while also increasing knowledge of how biologically and medically relevant Linked Data can be stored and queried while preserving the private nature of genomic data for Ghent University and VITO.

Within VITO, the PENGQUIN framework is to be designed to be integrated into ongoing research projects related to pharmacogenomics. More specifically, ongoing research investigating an improved system of drug prescription based on PGS factors for individual patients is currently underway at VITO. The aim for the PENGQUIN project is to provide a more complete, efficient, and reliable framework for patient genome storage and querying to improve the efficiency and efficacy of how pharmaceuticals are prescribed to patients.

Lastly, the algorithms and methodologies developed will be made available for future research. This way, the research will continuously contribute to academia and industry.

MORE HERE 🡪

At VITO:

pharmacogenomics projects

project “SAVE DATA”

Kinda at VITO:

WE ARE initiative in data privacy and protection

Belgian Genome Biobank

aspirations for Institutions partnered with VITO, JESSA and Ghent University such as Domus Medica along with other hospitals

*Elaborate the strategic dimension of your research, with regard to the (long-term) potential for innovative applications .*

*Substantiate the PhD project’s strategic focus on economically relevant innovations. Justify how the chosen research approach (if successful) is the appropriate one to achieve the anticipated application(s) (potentially long term).*

*Elaborate the strategic importance of the potential applications to possible users (impact). Show how (if the project is successful) new products, services and/or processes may affect business of specific companies, a collective of companies and/or a sector and/or may be closely aligned with the Flemish science, technology and innovation transition priorities* ([Flanders in transition. Priorities in Science, Technology and Innovation towards 2025](https://www.vlaanderen.be/publicaties/flanders-in-transition-priorities-in-science-technology-and-innovation-towards-2025)) (socio-economic benefits).*Societal impact should always be linked to a (in)direct (macro)economic benefit, e.g. cost reductions in health care, higher education level, environmental impact etc. should be positioned in an economic context.*

**5. References**

1. Souche, E. *et al.* Recommendations for whole genome sequencing in diagnostics for rare diseases. *Eur J Hum Genet* **30**, 1017–1021 (2022).

2. Gil, M. M., Quezada, M. S., Revello, R., Akolekar, R. & Nicolaides, K. H. Analysis of cell‐free DNA in maternal blood in screening for fetal aneuploidies: updated meta‐analysis. *Ultrasound in Obstet &amp; Gyne* **45**, 249–266 (2015).

3. Alzu’bi, A., Zhou, L. & Watzlaf, V. Personal genomic information management and personalized medicine: challenges, current solutions, and roles of HIM professionals. *Perspect Health Inf Manag* **11**, 1c (2014).

4. Tommel, J. *et al.* Personal Genomes in Practice: Exploring Citizen and Healthcare Professionals’ Perspectives on Personalized Genomic Medicine and Personal Health Data Spaces Using a Mixed-Methods Design. *Genes* **14**, 786 (2023).

5. Wang, L., McLeod, H. L. & Weinshilboum, R. M. Genomics and Drug Response. *N Engl J Med* **364**, 1144–1153 (2011).

6. Ko, Y. K. & Gim, J.-A. New Drug Development and Clinical Trial Design by Applying Genomic Information Management. *Pharmaceutics* **14**, 1539 (2022).

7. Hens, K., Nys, H., Cassiman, J.-J. & Dierickx, K. The return of individual research findings in paediatric genetic research. *Journal of Medical Ethics* **37**, 179–183 (2011).

8. McLeod, H. L. Cancer Pharmacogenomics: Early Promise, But Concerted Effort Needed. *Science* **339**, 1563–1566 (2013).

9. Zhu, X. *et al.* Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. *Genetics in Medicine* **17**, 774–781 (2015).

10. Marwaha, S., Knowles, J. W. & Ashley, E. A. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Med* **14**, 23 (2022).

11. Blumenthal, D. & Squires, D. Giving Patients Control of Their EHR Data. *J GEN INTERN MED* **30**, 42–43 (2015).

12. Damen, D. J. *et al.* Patients Managing Their Medical Data in Personal Electronic Health Records: Scoping Review. *J Med Internet Res* **24**, e37783 (2022).

13. Spagnuelo, D., Bartolini, C. & Lenzini, G. Qualifying and measuring transparency: A medical data system case study. *Computers & Security* **91**, 101717 (2020).

14. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*. *OJ* 1–88 (2016).

15. Piovesan, A. *et al.* On the length, weight and GC content of the human genome. *BMC Res Notes* **12**, 106 (2019).

16. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**, 1767–1771 (2010).

17. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

18. Bonfield, J. K. CRAM 3.1: advances in the CRAM file format. *Bioinformatics* **38**, 1497–1503 (2022).

19. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

20. Mansour, E. *et al.* A Demonstration of the Solid Platform for Social Web Applications. in *Proceedings of the 25th International Conference Companion on World Wide Web - WWW ’16 Companion* 223–226 (ACM Press, Montr&#233;al, Qu&#233;bec, Canada, 2016). doi:10.1145/2872518.2890529.

21. Umbrich, J., Hose, K., Karnstedt, M., Harth, A. & Polleres, A. Comparing data summaries for processing live queries over Linked Data. *World Wide Web* **14**, 495–544 (2011).

22. Home Page - Genomics Reporting Implementation Guide v3.0.0-ballot. http://hl7.org/fhir/uv/genomics-reporting/ImplementationGuide/hl7.fhir.uv.genomics-reporting.

23. Van Der Horst, E. *et al.* *Bridging Clinical and Genomic Knowledge: An Extension of the SPHN RDF Schema for Seamless Integration and FAIRification of Omics Data*. https://www.preprints.org/manuscript/202312.0373/v1 (2023) doi:10.20944/preprints202312.0373.v1.

24. Cyganiak, R., Wood, D. & Lanthaler, M. Concepts and Abstract Syntax. *W3C* (2014).

25. Berners-Lee, T. Linked Data. (2009).

26. Taelman, R. & Verborgh, R. Evaluation of Link Traversal Query Execution over Decentralized Environments with Structural Assumptions. (2023) doi:10.48550/ARXIV.2302.06933.

27. FitzHenry, F. *et al.* Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership. *Appl Clin Inform* **06**, 536–547 (2015).

28. Shin, S. J. *et al.* Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study. *J Med Internet Res* **21**, e13249 (2019).

29. Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R. & Stiawan, D. The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities. *JMIR Med Inform* **9**, e21929 (2021).

30. Fiume, M. *et al.* Federated discovery and sharing of genomic data using Beacons. *Nat Biotechnol* **37**, 220–224 (2019).

31. Rambla, J. *et al.* Beacon v2 and Beacon networks: A ‘lingua franca’ for federated data discovery in biomedical genomics, and beyond. *Hum Mutat* **43**, 791–799 (2022).