2nd ANNUAL REPORT OF THE NORDIC INFORMATION FOR ACTION eSCIENCE CENTER OF EXCELLENE (NIASC)

PROGRESS AND RESEARCH RESULTS DURING 2015

The overall goal of the Nordic Information for Action eScience Center (NIASC) is to use eScience-based tools and algorithms to enable introduction of risk-stratified cancer screening programs. This will enable the long awaited goal of moving from one-size-fits-all screening programs, to screening programs adapted to an individual person's risk. This will lead to both more effective and more cost-effective population-based screening programs. Our aim is to use eScience to exploit the best of the Nordic register and biobank-derived information, to advance screening programs also in a wider perspective. This is done by:

- Promoting Open Access to biobank samples and biomedical data.
- Improving information and communication technology (ICT), by developing and improving open source tools for data collection and informed consent (e.g. mobile apps).
- Increasing the understanding of the ethical challenges for effective data-mining of big biomedical data.
- Advance generic eScience tools for analysis of molecular and genetic data, for text mining of patient records and for safe and efficient management of highly variable datasets.
- Optimize risk prediction algorithms for screening, by utilizing the excellent Nordic registry and organized screening environment.
-

NIASC Key performance indicators

As requested by the Scientific Advisory Board (SAB) at the second annual meeting, NIASC has developed key performance indicators (KPIs) to objectively and efficiently quantify its success, performance and progress toward the main goal. The development of the KPIs is a deliverable of the work for 2015 and the current values of the Five KPIs are reported below:

Name	**Description**	Current number of KPI	Estimated total number of KPI by the end of project
KPI1	***Risk stratification strategies using NIASC eScience tools actually implemented in national cancer screening program(s) for risk stratified screening.***	0	2
KPI1a	*as above but in development or pilot implementation stage*	2	0
KPI2	***Number of Open Source eScience tools developed***	2	11
KPI2a	*as above but in development stage*	9	0
KPI3	***Number of research datasets where open access is promoted using NIASC eScience tools***	0	5
KPI3a	*as above but in development stage*	5	0
KPI4	***Number of joint international projects (>=3 Nordic countries)***	4	4
KPI5	***Academic Outputs*** *(Number of Papers published; Number of Workshops and Courses organized; Impact factor of the consortium)*	8	42

NIASC coordinating center activities

The NIASC coordinating center at the Karolinska Institutet is responsible for the overall management of the program. This includes tasks like keeping track of KPIs, budgets and finances as well as ensuring that the work of working groups and projects are proceeding well (or, if not, to take appropriate action).
Setting of the strategic agenda for NIASC and international benchmarking are other important tasks. Major strategic issues are presented to the SAB and to the consortium as a whole in the Annual Assembly meetings.
Finally, the coordinating center performs a number of development projects that are considered to be of key strategic relevance for the development of NIASC as a whole.

During 2015, major activities have been:

Strategic development projects

1) Development of new, open source IT systems for risk-stratified screening

We first mapped the IT systems used for cervical screening in the participating countries in the NIASC working group for cervical cancer. Only Norway used an entirely in-house-developed system (The other countries either used commercial software or a mix of commercial software and in-house developed functionalities). The national cervical screening program of Norway agreed to make their software Open Source and deposit it at the NIASC coordinating center.
In the autumn of 2015, we received the entire program code for cervical screening that is in routine use in Norway (developed in-house at the Cancer Registry of Norway, who deposited the program).
The coordinating center started with learning how the program works and translating program user interface (and the manual enclosed from the Cancer Registry of Norway) to English. We have then proceeded to modify the program to be suitable for use also in other screening services, both in Sweden as well as in other countries. This would provide a most valuable program commodity and eScience tool for the screening community both in the Nordic countries and beyond. This contributes to KPI 2a and when evaluated as KPI2 will have significant effect on KPI 1.
The vision is to develop a functionality that assigns risk by accessing available big biomedical data and automatically issues invitations to screening in an intensity that is tailored to the risk of the individual. We considered that such developments could be difficult and time-consuming if developed as add-ons to commercial software (which are proprietary and are also frequently changed as a result of purchasing) and that we needed to develop this as an add-on to an entirely free, open source screening system.

2). Development of a new Open Source system for quality assurance and improvement of screening.

The NIASC coordinating office has also received the source code of the NORDCAN application (http://www-dep.iarc.fr/nordcan.htm) from WHO/IARC (http://www-dep.iarc.fr). This Internet application enables an on-demand linkage to the national cancer registries in the Nordic countries to display aggregated data on incidence, mortality, prevalence and survival statistics of major cancers in the Nordic countries, with graphical and tabulation facilities. Based on NORDCAN, we are now developing a similar open-source application – "NORDSCREEN" - that will be able to on-demand retrieve key information on cancer screening (in particular key quality indicators) from

the Nordic Screening Registries and display aggregated data on a website. KPI 3 will measure success of this development as it will display available data in screening registries, as well as KPI2a as NORDSCEEN is under the development and will be available as open source software.

3.) Development of Open Source software for parallel analysis of metagenomic data

Next generation sequencing (NGS) is a most powerful source of very big biomedical data. Analysis of NGS data for human genomics can use the fact that the human sequence is already known for a rapid alignment, whereas bioinformatics for microorganisms (metagenomics) will need to match all reads against each other to accomplish assembly. Thus, metagenomics is very computer intensive requiring both use of parallel computing and powerful computing resources. At the same time, microorganisms are increasingly shown to be important determinants of health and disease. Therefore, it is envisaged that metagenomics will become a key source of data for risk prediction algorithms.

The NIASC coordinating office, in collaboration with professor Piotr Bala's group at the Polish Grid Infrastructure at the University of Warsaw, is developing a bioinformatics pipeline to analyze viral metagenomics datasets, generated by next generation sequencing (NGS) platforms. As part of this collaboration, Piotr Bala's group modified the NCBI BLAST source code to speed up the performance of the NCBI blastn algorithm. The modified blastn algorithm PCJ-BLAST is now 10 times faster on the PLGRID infrastructure and is available through the PLGRID resources. Currently, NIASC coordinating office is able to, with remote access, utilize computing resources of the Polish Grid Infrastructure (http://www.plgrid.pl) and run PCJ-BLAST algorithm on 32 computing nodes with 28 CPUs each (896 CPUs total). Later, the workflow will also have a graphical user interface and will provide a valuable eScience tool for the microbiome research communities in Nordics and beyond. Both the workflow and the modified blastn algorithm will be available as open source. Thus, this development is a KPI 2a.

4.) Best Practices for Big Data Handling in Biobanking

The advances in biomedicine and related information and communication technologies result in that waves of huge and complex health-related data, also referred to as "Big Biomedical Data" is arriving. Biobanks are increasingly becoming big data banks, rather than sample repositories. However, storing, analyzing and sharing this Big Biomedical Data and converting it to actionable knowledge have technical challenges. Handling Big Biomedical Data in biobanks was defined as one of the strategic topic by NIASC IT/Bioinformatics group and by the coordinating office. In November 2015, NIASC organized a workshop that brought together Nordic national biobanks, national biobank infrastructures and Nordic e-infrastructure service providers for sensitive data processing. The main purpose of the meeting was to exchange knowledge and experiences how to build and use efficient and sustainable e-infrastructures, as well as setting up and maintaining associated ecosystem of workflows, pipelines and bioinformatics software for storing, analyzing and sharing large amounts of genomic and other health-related data. During the workshop, we found that the community of Nordic national biobanks, national biobank infrastructures and e-infrastructure service providers have many common interests and there is much to gain from engaging in collaborations in these areas of common interest. Three working groups were formed that will i) formulate a joint use case on IT needs for the Nordic National Biobanks & Infrastructures. ii) develop and optimize useful bioinformatics software/algorithms for metagenomics that are scalable and able to efficiently utilize parallel computing and iii) develop IT solutions that will enable mapping of multiple outcomes from registries to different biological entities (Reverse data mining or "registromics"). The workshop itself contributed to KPI 5. Established working groups have great potential to further enhance collaboration of Nordic countries and development of eScience tools and open source software - impacting on all 5 KPIs

NIASC coordinating center performance evaluated by KPIs

Name	Description	Current number of KPI	Estimated number of KPI by the end of the project
KPI1	***Risk stratification strategies using NIASC eScience tools actually implemented in national cancer screening program(s) for risk stratified screening.***	0	1
KPI1a	*as above but in development or pilot implementation stage*	1	0
KPI2	***Number of Open Source eScience tools developed***	0	3
KPI2a	*as above but in development stage*	3	0
KPI3	***Number of research datasets where open access is promoted using NIASC eScience tools***	0	1
KPI3a	*as above but in development stage*	1	0
KPI4	***Number of joint international projects***	0	0
KPI5	***Academic Outputs*** *(Number of Papers published; Number of Workshops and Courses organized; Impact factor of the consortium)*	1	10

International Collaborations initiated by the Coordinating Office

The establishment of NIASC has aroused significant interest among several major partners in the cancer epidemiology field. We have thus made several key contacts, all in the spirit of open access, to collaborate in the cancer prevention field:

a) As a direct result of contacts made within NIASC, the NIASC partners ***Icelandic Cancer Society*** and Karolinska Institutet/Karolinska University Hospital have entered into a unique collaboration. Iceland sends the samples from their cervical screening program for human papillomavirus (HPV) testing by the Karolinska Institutet in Sweden, instead of establishing their own physical lab at large costs and efforts. Correct delivery routines, and IT systems capable of handling transnational sample reports were developed at the NIASC coordinating office. The testing is live as of fall 2015. As far as we are aware, it is unique for an organized screening program to send the samples for testing abroad. Possibly, the system used could facilitate the structural development also of other screening programs and has potential to have an important Nordic added value contributed from NIASC.

b) ***The National Cancer Institute at the National Institutes of Health (NCI/NIH)***, Washington DC, USA. Having presented NIASC and our main goals for leading cancer researchers (Hormuzd Katki and Mark Schiffman), the Division Director Stephen Chanock has officially approved a collaboration between NCI and NIASC. The initial collaboration will center on the open access sharing of large-scale data (with on-line access for the US researchers to all data in a national screening registry as a use case), in order to further scientific advances and collaboration across borders. We are also discussing collaboration on the development of mobile phone applications (for risk prediction and also gathering of informed consent).

c) **EGI Engage** – The project aims to create a pan-European e-infrastructure for biobank data sharing and analysis. NIASC participates with a use case: a bioinformatics pipeline to analyze viral metagenomics datasets (developed together with University of Warsaw – see strategic projects above). This is considered as a strategic initiative to further dissemination of achievements.

d) **EUDAT – European data infrastructure** - aims to contribute to the production of a Collaborative Data Infrastructure and provide a pan-European solution to the challenge of data proliferation in Europe's scientific and research communities. NIASC has filed an application to use EUDAT infrastructure for a medical science use case project that uses machine-learning techniques for cervical cancer risk prediction (input data from Swedish National Quality Register for Cervical Cancer Prevention. Led by professor Jan Komorowski, see description of projects below).

e) **B3Africa –** EU funded project under the Horizon 2020 work programme. B3Africa aims to Bridge Biobanking and Biomedical research across Europe and Africa. Eleven partners from African and European countries are participating to develop a collaboration framework and an open source informatics infrastructure that will accelerate and facilitate biomedical research across the continents to address global health challenges together. KI is a B3Africa partner and is leading the work on defining prioritized use cases that will need development of open source software and affordable informatics platforms.

NIASC Recruitments

NIASC has recruited a total of 13 positions, all of which have been announced at the home department's web page, on our NIASC web page (www.nordicehealth.se) and via the EURAXESS portal, to promote open international competition. The gender distribution of new positions in NIASC is as follows, based on recruitments for coordination purposes, DBA's/Bioinformaticians, and the positions granted in the first internal call:

Position	Gender	Name
Research Coordinator (S)	Female	Karin Sundström
Bioinformatician (S)	Male	Davit Bzhalava
PhD-student (S)	Male	Nicholas Baltzer
PhD-student (S)	Female	Rebecka Weegar
Postdoc (F)	Male	Kimmo Palin
Bioinformatician (D)	Male	Victor Yasimov
PhD-student (IS)	Female	Thorgerdur Palsdottir
Database manager (F)	Male	Kimmo Pääkkonen
Postdoc (D)	Female	Xueping Liu
Bioinformatician (N)	Male	Sigurd Gartmann
Postdoc (F)	Male	Priit Palta
Midwife (S)	Female	Agneta Carlsten Thor
Postdoc (N)	Male	Tomás Ruiz-López

(S) = Sweden, (N) = Norway, (D) = Denmark, (F) = Finland, (IS) = Iceland

In total: this means that the gender balance of recruitments from ads for new positions is currently 5 females and 8 males. The balance currently comes to 5:8 in terms of definitive positions, which we regard as a highly successful outcome given the dominance of male candidates in some eScience-related disciplines such as IT and computer science. This gender distribution is also in line with the declaration in our original application to promote gender equality for younger researchers.

3.1 NIASC IT/bioinformatician/DBA team

Four IT/bioinformatician/DBAs were recruited in Norway, Sweden, Denmark and Finland, respectively. In January 2015, during the NIASC 2nd annual meeting, the NIASC IT/bioinformatician/DBA working group met for the first time. In March 2015, the group had a meeting focusing on to get to know each other, to identify the most critical aspects for collaboration and to plan future activities. This working group is very strong in different programming languages, machine learning and statistics. Expertise of this group will be key for further cross-border, NIASC-internal collaborations. Building scalable infrastructure for storing, analyzing and sharing big biomedical data across the Nordic countries, as well as the pilot of the National Biobank Registry System was identified as highly important topics for the group to work and collaborate on. The team will ensure software sustainability and will follow the guidelines developed by software sustainability institute (http://www.software.ac.uk/software-evaluation-guide).

Dissemination of results

NIASC has twice featured prominently in the NordForsk magazine. The NIASC website (www.nordicehealth.se) has been updated to promote dissemination of results and resources. For 2016, we plan on launching a webinar series, an internet-based educational platform and a Newsletter.

Research mobility

Please specify research stay abroad as well as visits by foreign researchers. Here mobility is defined as a stay abroad of at least 2 weeks duration.

Name, job title, organisation	Site of work	Purpose of visit	Duration of visit	Comments, output of the visit
Thorgerdur Palsdottir, PhD student, KI	Stockholm, Sweden	PhD Studies	4 years	
Priit Palta, PostDoc, FIMM	Helsinki, Finland	Postdoc research	2 years	
Total number of visiting months: 6 years				
Total number of visiting researchers: 2				

As a result of NIASC activities and communication between the partners during the 2015 2 main long-term researches mobility was observed:

1. Thorgerdur Palsdottir from Laufey Tryggvadóttir's group (Icelandic Cancer Registry) moved to Karolinska Institutet at Mark Clements group as a doctoral student.
2. Priit Palta from Andres Metspalu's group (Estonian Genome Center) moved as postdoctoral researcher to Institute for Molecular Medicine Finland (FIMM) at Aarno Palotie's group.

Meetings and Networking

During 2015 NIASC arranged several networking meeting:

1. As described above in the activities of NIASC IT/Bioinformatician/DBA team, in March 2015 the group organized a meeting focusing on to get to know each other, to identify the most critical aspects for collaboration of the group and to plan future activities. As a direct

result of this meeting we organized a workshop about Best Practices for Big Data Handling in Biobanking (described on page 5).

2. As described on page 5 under the heading "Best Practices for Big Data Handling in Biobanking" in November 2015, NIASC organized a workshop that brought together Nordic national biobanks, national biobank infrastructures and Nordic e-infrastructure service providers for sensitive data processing. During the meeting, participants exchanged knowledge and experiences of how to build and use efficient and sustainable e-infrastructures, as well as setting up and maintaining associated ecosystem of workflows, pipelines and bioinformatics software for storing, analyzing and sharing large amounts of genomic and other health-related data.
3. NIASC cervical cancer working group held its 2nd annual meeting on 16th of June, 2015 in Stockholm, Sweden. During the meeting participants presented progress reports on ongoing NIASC cervical cancer projects, also a round-table discussion about joint Nordic projects on cervical cancer research was organized.

External funding

We have applied for 5M€ from the EU Horizon2020 call on Big Data (EBBA- European Big Biomedical data for Action). No external funding received as yet.

NIASC Work plan for 2016 and onwards

To achieve our main goal: The real-life introduction of risk-stratified organized cancer screening programs, we will continue to follow our 2 main tracks:

1) Development of best risk prediction markers for prostate, cervical and colorectal cancer screening practices, as well as development of methods to advance screening engagement of citizens in cancer screening programs.

2) Development of IT-systems and software capable of handling the introduction and management of risk-stratified cancer screening programs.

The tasks of the Center Coordinating Office, the Working Groups, the PhD projects, the PostDoc projects and the national coordinating DBA/Bioinformaticians are clear and will not change during 2016.

During the first 2 years, a focus of the Coordinating Office has been to ensure that the international PhD projects and PostDoc projects are designed, evaluated and implemented and that the recruitments of national coordinating DBA/Bioinformaticians proceed well. As these tasks are by now all completed, with all PhDs, PostDocs and DBA/Bioinformaticians recruited, we will during 2016 focus on developing **Integrating Activities** to ensure cross-project communication, Nordic added value and the creation of a "NIASC community".

Modes of action for this that will be piloted are:

a) Regular **Webinars.** The contemplated format is webinars with 2 x 20 minutes presentations given by NIASC employees (with webinars rotating between countries). The intention is that participation in the Webinar series should be considered a compulsory part of the scientific education in NIASC. The target group is other NIASC participants, but they will also be open to external participants.

b) Internet-based education. We will expand key webinar presentations, in particular those concerning best practices and services, to internet-based educational courses. These are primarily intended for the NIASC participants, but will be Open Resources. We will use the Open Source educational platform Moodle (https://moodle.org), that enables posting both

slides, reading materials, videos and exams. A course testing the system has been posted at booc.biobanks.se

c) Increased activity from the coordinating office to promote the use of NIASC funds for **bilateral exchange.** A substantial part of the NIASC budget is reserved for bilateral exchange for personnel from one Nordic country to visit another Nordic country.

d) NIASC coordinating office will regularly **measure performance based on the KPIs** described in this report and increasingly encourage the different working groups to meet physically, with representatives also from the Coordinating Office present.

e) **Improve the presence of NIASC online.** This involves improving the website and keeping it updated, issuing of a NIASC Newsletter and drafting of an explicit, written communication and dissemination plan.

Deviations from original plan and possible budget implications

All key aspects of the proposal (Coordinating DBAs/Bioinformaticians hired and working together; Internal calls for PhD and PostDoc positions to work on different aspects of the NIASC program) were completed as planned. Having completed the building of the structure of the consortium, NIASC will now be in a stable phase with focus on completing the proposed work.

Delays in recruiting personnel were limited, but have still resulted in an underspending of about 400000 Euros. As this is a one-time available sum, we cannot use it for additional positions but will during 2016 be launching so-called **"seed projects".** These are defined as projects that have a defined start and end, but could still be expected to result in a sustainable, long-lasting impact that is in line with NIASC overarching goals. Several such projects have been proposed by the partners and will be implemented during the fall of 2016.

Several **emerging opportunities** have been added to the original plan (EGI-Engage, EUDAT, NCI, IARC, NORDCAN, Iceland-Sweden collaboration on HPV screening), but are not expected to incur any significant extra costs as each emerging opportunity is responsible for its own costs and mostly receives support from NIASC in terms of sharing of available resources.

NIASC considers that the resources and results are key components for addressing the challenge of how best to exploit **Big Biomedical Data.** Several funding agencies, notably Horizon 2020 and several Nordic funding agencies have announced that exploiting Big Biomedical Data is a current key priority. The NIASC coordinating office will be monitoring the development of this field and associated opportunities that may arise for contributions by NIASC to this field.

NIASC Research Projects and main results including publications

For the progress made during our 1st and 2nd year, we here refer to the plan developed in our original application, S*ection B: Development of eScience tools* and *C. Pilot Projects*.

B1. Nordic Biobank Registers

Project Title	*Nordic Biobank Registers*
Project Leader	Mads Melbye
Institution(a)	Statens Serum Institut, NTNU
Date started	2015-01-01

Estimated completion date	2017-12-31

LIST OF PERSONNEL EMPLOYED BY THE PROJECT

Name	Position	Academic Degree	Gender
Xueping Liu	PostDoc	PhD	Female

CURRENT STATUS

Report on progress

The launch of Nordic Biobank Registers (NBR) would be one of the major achievements by NIASC that will support open access to available biobanked samples and other health data registers. The mainstay for this development is refining and extending the concept already established at the Danish National Biobank, which hosts an open access, free-of-charge online web portal system for linking biobanked samples to health data registers, thus displaying which samples with associated health & disease data that are available for potential new research projects (please see www.biobanks.dk). In April 2014, a postdoc project dedicated to this development was approved for funding in the first internal call (main supervisor professor Mads Melbye). The postdoc position was filled as of December 2014. The Danish NIASC node implemented a prototype in Scala programming language and using Akka framework (see diagram below), which is a first step to adaptation of the current Danish Biobank Register system to serve also in other countries. This new model allows an asynchronous management of requests and flexibility in adding multiple databases to the system, thus multiple countries can become integrated all together. On top of that, there is a logging database (called NBR DB) designed, in order to track all the requests and responses of the system.

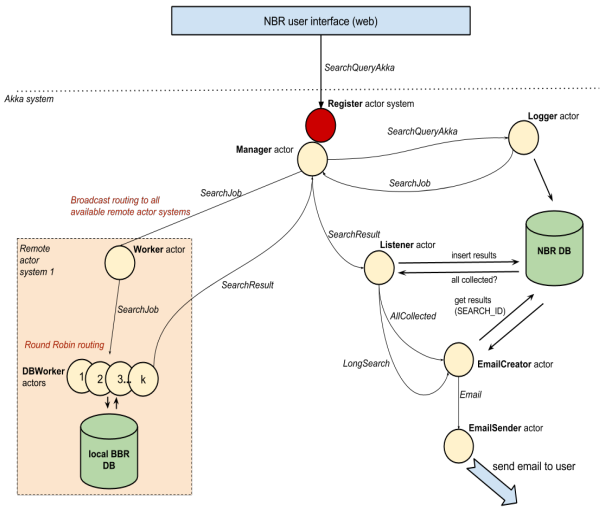
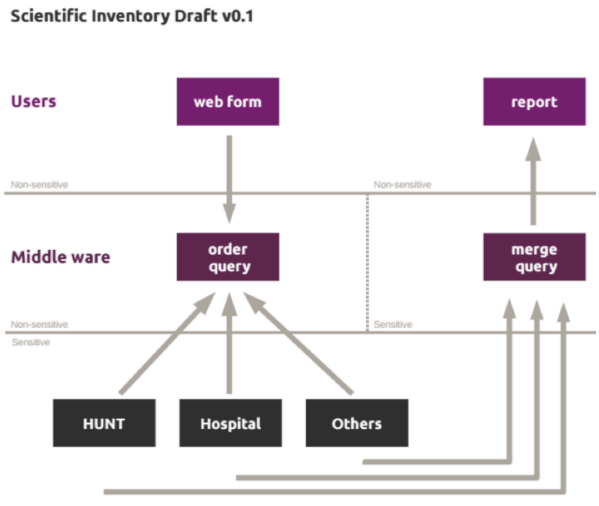


Figure 2. In the Norwegian node a prototype of registry lookup service is implemented, using synthetic data with generated correlations. The prototype is implemented as multiple microservices using the Python programming language, the "socketservice" module for communication and "gnuPG" for channel encryption. Pseudo-IDs are encrypted using "scrypt". An algorithm was designed to make it costly to perform hardware attacks. As a next step a pilot implementation of the prototype is planned. Three Norwegian biobanks are willing to participate in the pilot (HUNT, Mor-Barn and Tromsø). Permission from the Norwegian data protection agency will be obtained to install the registry lookup service



at pilot registries/biobanks. Also a searchable GWAS catalogue for HUNT study variables has been made and it is planned to incorporate it in the Nordic Biobank Register.
NIASC has used the software from the Danish National Biobank Register, www.biobanks.dk, and installed at NIASCs Coordinating Office at Karolinska Institutet, Stockholm, Sweden. www.biobanks.se as a concept has been approved by the Swedish Data Inspection and ethical approval was obtained to use this as a platform for a register-based linkage service that will increase the awareness of, and simplify access to, samples complete with relevant enriching information on outcomes. As a first pilot implementation, available biobanked samples from the large-scale STHLM2 study (a cohort study on prostate cancer detailed in our original application.) is now displayed on www.biobanks.se;

PROJECT PERFORMANCE

Estimate project according NIASC Key Performance Indicators (KPIs)

Name	**Description**	Current number of KPI	Estimated number of KPI by the end of the project
KPI1	***Risk stratification strategies using NIASC eScience tools actually implemented in national cancer screening program(s) for risk stratified screening.***	0	0
KPI1a	*as above but in development or pilot implementation stage*	0	0
KPI2	***Number of Open Source eScience tools developed***	0	0
KPI2a	*as above but in development stage*	0	0
KPI3	***Number of research datasets where open access is promoted using NIASC eScience tools***	0	1
KPI3a	*as above but in development stage*	1	0
KPI4	***Number joint international projects***	1	1
KPI5	***Academic Outputs*** *(Number of Papers published; Number of Workshops and Courses organized; Impact factor of the consortium)*	0	2

Shortly describe the contribution of the project to each KPIs (from column "Current number of KPI")

The success of NBR will be a NIASC KPI3, currently in development or pilot implementation phases, therefore contributing to NIASC KPI3a. As this project currently involves collaboration of 3 Nordic countries (Denmark, Norway and Sweden) it also contributes to KPI 4.

CURRENT FINANCES

Estimated total cost	1 624 978 NOK
Total expenditure by 2015-12-31	1 099 287 NOK

Planned expenditure by 2015-12-31	1 099 287 NOK

B2. E-science tools for handling and validation of next-generation sequencing data and B3. Nordic standards for identification of incidental findings of genetic risk and feedback to citizens

Project Title	Developing an efficient imputation pipeline to construct near complete genome variant data information in GWAs datasets
Project Leader	Aarno Palotie
Institution(s)	Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Estonian Genome Center (EGC), University of Tartu
Date started	01.09.2015
Estimated completion date	31.08.2017

LIST OF PERSONNEL EMPLOYED BY THE PROJECT

Name	Position	Academic Degree	Gender
Priit Palta	PostDoc	PhD	Male

CURRENT STATUS

Report on progress

We have been working on improving the 1000 Genomes Project -based imputation pipeline by using Finnish low-coverage (~4.6x) whole genome sequencing data. Figure 1 represents the improvement of rare and low-frequency variant imputation accuracy when local population specific data are included in the imputation reference panel.

By using near complete genome variant data for 20 000 Finns (originally genotyped on different GWAS chips and now imputed into our current low-coverage imputation panel), we are studying if variants enriched in Finland are associated to different quantitate traits (of cardiovascular and immunological relevance) measured in these individuals and/or different disease endpoints defined from their health registry data.

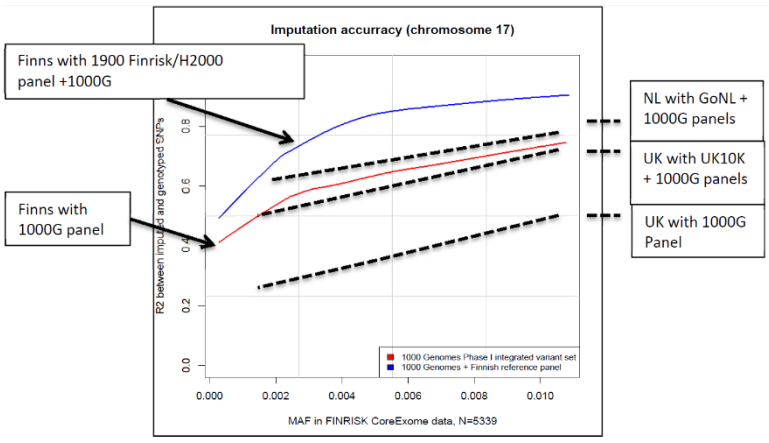


Figure 1. Imputation accuracy of SNP with MAF <= 1% in FINRISK HumanCoreExome sample set using two different imputation reference panels. Singletons, doubletons and SNPs with "info" < 0.4 were filtered out.

Additionally, we have started working on harmonization of the variant data and on constructing the improved imputation pipeline that additionally to the publicly available reference datasets (eg. 1000 Genomes Project) would take advantage of the high-coverage whole genome sequencing datasets currently available for several thousands of Finns and for ~2000 Estonians.

Updated plan (if applicable)

In close collaboration with the Estonian Genome Center we have now completed the construction of the new Estonian imputation reference panel utilising 2300 high-coverage whole-genome sequenced (hcWGS) Estonians. This data resource will be now used for variant imputation followed by the consequencive downstream association analyses.

We are currently working on a similar hcWGS dataset including 2200 Finns. This data will be put through our rigorous QC steps and then statistically phased to construct the new high-coverage Finnish imputation reference panel.

PROJECT PERFORMANCE

Estimate project according NIASC Key Performance Indicators (KPIs)

Name	**Description**	Current number of KPI	Estimated number of KPI by the end of the project
KPI1	***Risk stratification strategies using NIASC eScience tools actually implemented in national cancer screening program(s) for risk stratified screening.***	0	0
KPI1a	*as above but in development or pilot implementation stage*	0	0
KPI2	***Number of Open Source eScience tools developed***	0	1
KPI2a	*as above but in development stage*	1	-
KPI3	***Number of research datasets where open access is promoted using NIASC eScience tools***	0	1
KPI3a	*as above but in development stage*	1	-
KPI4	***Number joint international projects***	2	2
KPI5	***Academic Outputs*** *(Number of Papers published; Number of Workshops and Courses organized; Impact factor of the consortium)*	0	2

Shortly describe the contribution of the project to each KPIs (from column "Current number of KPI")

The results of this project will be measured by KPI2, but because it is in the development stage it is now measured as KPI2a. It should be noted that datasets of this project will be made available to other researchers, therefore it will also contribute to KPI3. * In collaboration with the Estonian Genome

Center we are preparing a technical comparison paper using the Estonian imputation reference panel ('Improved imputation quality of rare and low-frequency variants using Estonian population-specific imputation reference panel').

CURRENT FINANCES

Estimated total cost	NOK 1 500 000 (EUR 187 046)
Total expenditure by 2015-12-31	NOK 222 683 (as of 31.12.2015)
Planned expenditure by 2015-12-31	NOK 758 796 (underspent due to late recruitment, which carryforward was discussed and approved by the coordinator)

B3. Computational methods for genetic cancer susceptibility analysis

Project Title	*Computational methods for genetic cancer susceptibility analysis*
Project Leader	Lauri Aaltonen
Institution	University of Helsinki, Norwegian University of Science and Technology
Date started	2014-09-01
Estimated completion date	2018-08-31

LIST OF PERSONNEL EMPLOYED BY THE PROJECT

Name	Position	Academic Degree	Gender
Kimmo Palin	Post Doc	PhD	Male

CURRENT STATUS

Report on progress

During the first year of the project, methods for analysis of both somatic and germline cancer genomes were developed and applied. An article "CTCF/cohesin binding sites are frequently mutated in cancer" was published in Nature Genetics (Nature Genetics 47, 818-821, 2015) where the novel phenomenon that CTCF/cohesin binding sites in the non-coding genomic DNA are frequently mutated in certain tumors are described. These tumors are characterized by a peculiar mutational signature, where TT dinucleotides tend to mutate to GT (And AA to AC in the opposite strand). This spatial-global association was first discovered in whole genome sequenced colorectal cancer tumors and further validated in several other tumor types that have been sequenced in large international projects. Causes and consequences of the discovered phenomena remain unknown. In parallel to the somatic analysis, 1700 colorectal cancer patients were genotyped on Illumina SNP arrays and 7,000 control samples through collaboration with Finnish Institute of Health and Welfare and Finnish Institute of Molecular Medicine. As the limited case series did not reveal novel associations in the national GWAS approach, the project joined forces with COGENT consortium in UK to execute a meta-analysis of 13,656 cases

and 21,667 controls of European ancestry. This analysis revealed a novel CRC associated locus and provides new insights of relationship of CRC with inflammatory bowel disease. The plan for the future is to analyse tumor samples in more detail, focusing on the structural variation in the tumors, as well as to develop methods for novel DNA sequencing methods such as Oxford Nanopore.

PROJECT PERFORMANCE

Estimate project according NIASC Key Performance Indicators (KPIs)

Name	**Description**	Current number of KPI	Estimated number of KPI by the end of the project
KPI1	***Risk stratification strategies using NIASC eScience tools actually implemented in national cancer screening program(s) for risk stratified screening.***	0	0
KPI1a	*as above but in development or pilot implementation stage*	0	-
KPI2	***Number of Open Source eScience tools developed***	0	1
KPI2a	*as above but in development stage*	1	-
KPI3	***Number of research datasets where open access is promoted using NIASC eScience tools***	0	1
KPI3a	*as above but in development stage*	1	-
KPI4	***Number of countries involved in joint international projects and number of such projects***	0	0
KPI5	***Academic Outputs*** *(Number of Papers published; Number of Workshops and Courses organized; Impact factor of the consortium)*	1	4

Shortly describe the contribution of the project to each KPIs (from column "Current number of KPI")

The results of these projects will be measured by KPI2, but because it is in the development stage it is now measured as KPI2a. Also the publication (Nature Genetics 47, 818-821, 2015) contributed to KPI5. It should be noted that datasets of this project will be made available for other researchers, therefore it will also contribute to KPI3.

CURRENT FINANCES

Estimated total cost	2 712 417,48 NOK
Total expenditure by 2015-12-31	698 365 NOK
Planned expenditure by 2015-12-31	698 365 NOK

B4. Generic eScience tools and techniques for predictive models and micro-simulation

Project Title	*Machine learning methods for cancer risk stratification*
Project Leader	Jan Komorowski
Institution(s)	Uppsala University, Karolinka Institutet, Kreftregisteret
Date started	2014-08-01
Estimated completion date	2018-08-31

LIST OF PERSONNEL EMPLOYED BY THE PROJECT

Name	Position	Academic Degree	Gender
Nicholas Baltzer	PhD Student	MSc	Male

CURRENT STATUS

Report on progress

A PhD-student trained in mathematics, computer science and bioinformatics has been given access to the full dataset from the Swedish National Quality Register for Cervical Cancer Prevention, in order to run machine-learning techniques for cervical cancer risk prediction. The project was launched in January 2014 and aims to create a tool that can calculate a risk index based on a woman's entire past screening history, instead of just the current practice which only uses the most recent smear result. Classifier has been built and used to find variables that could separate cases and controls on a Swedish dataset. The overall accuracy was at 64% (AUC at 71%), with significantly better performance for subgroups. Some very high risk subgroups (OR 32.1) comprising almost 10% of the cases, and multiple medium-to-high risk groups (ORs 4-15) were identified. Of the lower than average risk groups, the lowest had OR 0.55. The largest subgroup (at around 72% of the entire study population) was asymptomatic using developed methods, and removing it increased classifier accuracy to 74% (AUC at 76%). The mathematical models are robust, and yield similar risk profiles regardless of timeframes and age of data. Findings have presented preliminary findings at HPV2015 in Lisbon. The classifier will be tested for transferability in Norwegian data. The intention is to make a validated and rational software openly available, for the use in organized screening programs.

PROJECT PERFORMANCE

Estimate project according NIASC Key Performance Indicators (KPIs)

Name	Description	Current number of KPI	Estimated number of KPI by the end of the project
KPI1	***Risk stratification strategies using NIASC eScience tools actually implemented in national cancer screening program(s) for risk stratified screening.***	0	0
KPI1a	*as above but in development or pilot implementation stage*	0	-

KPI2	***Number of Open Source eScience tools developed***	0	1
KPI2a	*as above but in development stage*	1	0
KPI3	***Number of research datasets where open access is promoted using NIASC eScience tools***	0	0
KPI3a	*as above but in development stage*	0	-
KPI4	***Number joint international projects***	0	0
KPI5	***Academic Outputs*** *(Number of Papers published; Number of Workshops and Courses organized; Impact factor of the consortium)*	0	4

Shortly describe the contribution of the project to each KPIs (from column "Current number of KPI")

Success of this project will be measured by KPI2. However, as they are in the development stage it is now measured as KPI2a.

CURRENT FINANCES

Estimated total cost	2 602 228 NOK
Total expenditure by 2015-12-31	890 881SEK
Planned expenditure by 2015-12-31	890 881SEK

B5. Prediction and microsimulation of prostate cancer in the Nordic countries

Project Title	*Prediction of Prostate Cancer Risk to Improve PSA Screening: A Statistical Modeling Approach*
Project Leader	Mark Clements
Institution(s)	Karolinska Institutet, Icelandic Cancer Society
Date started	01.05.2015
Estimated completion date	01.05.2019

LIST OF PERSONNEL EMPLOYED BY THE PROJECT

Name	**Position**	**Academic Degree**	**Gender**
Thorgerdur Palsdottir	PhD stutent	MSc	Female

CURRENT STATUS
Report on progress

The PhD position has been filled. As an example of Nordic mobility, the recruited Icelandic statistician moved to Stockholm. Extensive work has been conducted on standardisation and structuring of the STHLM0 PSA Register, which forms the basis for future publications for this project. Analysis has been conducted and a manuscript entitled "Outcomes of Prostate Cancer Screening by PSA Testing Frequency Summary" is in preparation. In outline, the objective of the investigation is to make PSA testing more personalized by comparing the harms and benefits of PSA testing frequency according to age, PSA levels and family history of prostate cancer. According to the preliminary results, total PSA, age and less frequent testing intervals are significantly associated with increased odds of being diagnosed with advanced prostate cancer. Analysis is currently ongoing for another manuscript entitled "Analysis Plan for Comparing a New Biomarker Test to the Conventional PSA Test". In outline, the objective of the investigation is to identify new biomarker tests that would improve the true detection rate and decrease the false detection rate and thus help to decrease mortality and over-diagnosis of the disease. Diagnostic performance of a new biomarker test called S3M and its performance relative to the PSA test is being evaluated in the Stockholm 3 study. This project will focus on the development of risk prediction methods for prostate cancer but the methods developed here, as for the Komorowski project, can be adapted for other cancer forms. We have also made some progress with (a) the prostate cancer natural history model for the microsimulation project and (b) the development of a mobile app for risk prediction. Below is more detailed description of these projects:

Outcomes of Prostate Cancer Screening by PSA Testing Frequency **Importance:** Controversy exists about the frequency that men should undergo PSA testing and whether testing intervals should vary according to risk factors. To find the optimal PSA testing frequency, the questions need to be answered whether less frequent screening results in a more advanced cancer for men diagnosed with prostate cancer and whether more frequent screening leads to more unnecessary, and possibly harmful, biopsies for men without cancer. **Objective:** We aim to make PSA testing more personalised by comparing the harms and benefits of PSA testing frequency according to age, PSA levels and family history of prostate cancer. **Data:** Data were collected from the STHLM0 PSA Register from 2008-2015 for all men aged 50-74 years that had a PSA test in Stockholm county. The total number of men that were tested for PSA was 207,120, of which 15,118 men were diagnosed with prostate cancer.**Main outcomes and methods:** We calculated the odds of advanced prostate cancer (defined as a cancer with a Gleason score of 7 or higher) according to screening frequencies, age, PSA levels and family history. We also calculated the odds of the cumulative probability of being referred to biopsy and not diagnosed with cancer after 1 or more PSA tests according to testing frequencies, PSA values, age and family history. **Preliminary Results and conclusions:** According to the prelimary results, the total PSA, age, less frequent testing intervals are significantly associated with a slightly increased odds of being diagnosed with advanced prostate cancer. This provides important evidence for assessing the benefits and harms of PSA testing, and suggests that less frequent testing may be associated with more advanced cancer diagnosis. As a potential limitation, we do not have sufficient power to assess whether these associations hold for men with PSA values below 1 ng/ml.

Analysis Plan for Comparing a New Biomarker Test to the Conventional PSA Test

New biomarker tests to screen men for prostate cancer are in great demand. Tests used in combination with existing screening tests that would improve the true detection rate and decrease the false detection rate would help to decrease mortality and over-diagnosis of the disease. The only established screening test for prostate cancer is the PSA test, which has acceptable true positive rate but the false positive rate of the test is considered too high. Thus, too may patients are experiencing unnecessary

biopsies and over-diagnosis of the disease that can cause harm for the patient and create an economic burden for the health system. This could be alleviated with better screening tests. In this analysis we develop methods to evaluate the diagnostic performance of a new biomarker test called SM3 and its performance relative to the PSA test. The S3M test was tested in a study in Stockholm county called the STHLM3 study (Grönberg et al 2015). These methods rely on the relative true and false positive rates to measure the loss in sensitivity and gain in specificity associated with the combination of the two tests relative to the performance of the PSA test only. We will also use a regression models fit to data from the Stockholm 3 study to evaluate the diagnostic performance of the test compared to the conventional PSA test conditional on covariates (e.g. age). The theory for performing regression models to paired screen positive designs was developed by Pepe and Alonzo 2001. They use marginal models fitted to all of the data, including those with missing disease status. However, in Stockholm 3, we do not have complete data on all individuals in the study, which is required in the framework due to Pepe and Alonzo. We propose an extension to the framework developed by Pepe and Alonzo, where the model is fitted to those individuals with known disease status. We have developed theory to show that the results from the regression analysis still generate the same results with complete data only on the men with verified disease status. A manuscript describing these results is ongoing.

Prostate cancer natural history project

In collaboration with Andreas Karlsson, we have made significant progress with the Nordic prostate cancer natural history model. This model is being developed using Swedish data and will then be applied to the Icelandic setting, supplemented with Icelandic PSA test patterns. The natural history model is now informed by Swedish PSA test patterns and Gleason-specific patterns of treatment and diagnosis (using the STHLM0 database), and survival has been calibrated to the Swedish National Prostate Cancer Register. The R and C++ code for the Nordic natural history model is available from https://github.com/mclements/microsimulation/tree/develop. Furthermore, we have investigated gaining access to the PSA data from the Icelandic Cancer Society. We are in the process of applying for ethical permission for these data.

Development of a mobile app for prostate cancer risk prediction

We have begun to investigate approaches to developing a mobile app for calculating the risk of prostate cancer. Our initial interest was in developing an app for clinicians that would be useful with the S3M test developed by the STHLM3 study. However, there are concerns with intellectual property with the S3M test, so we propose investigating risk prediction for either PSA testing or for all cause mortality for determining whether a man should undertake prostate cancer screening.

References

Grönberg H, Adolfsson J, Aly M, Nordström T, Wiklund P, Brandberg Y, Thompson J, Wiklund F, Lindberg J, Clements M, Egevad L, Eklund M. Prostate cancer screening in men aged 50–69 years (STHLM3): a prospective population-based diagnostic study. The Lancet Oncology. 2015 Dec 31;16(16):1667-76.

Significant changes to plan that have occurred (if applicable)

As described above, there may be some limitations on the development of the mobile app due to intellectual property. We are investigating other risk calculations that would be useful in such an app.

Updated plan (if applicable)

At present, there are no specific changes to the plan.

PROJECT PERFORMANCE

Estimate project according NIASC Key Performanc Indicators (KPIs)

Name	**Description**	Current number of KPI	Estimated number of KPI by the end of the project
KPI 1	***Risk stratification strategies using NIASC eScience tools actually implemented in national cancer screening program(s) for risk stratified screening.***	*0*	*0*
KPI 1a	*as above but in development or pilot implementation stage*	*0*	*0*
KPI 2	***Number of Open Source eScience tools developed***	*0*	*2*
KPI 2a	*as above but in development stage*	*2*	*-*
KPI 3	***Number of research datasets where open access is promoted using NIASC eScience tools***	*0*	*1*
KPI 3a	*as above but in development stage*	*1*	*0*
KPI 4	***Number of countries involved in joint international projects (>=3) and number of such projects***	*0*	*0*
KPI 5	***Academic Outputs*** *(Number of Papers published; Number of Workshops and Courses organized; Impact factor of the consortium)*	*0*	*Number of papers: 5* *Number of workshops: 1*

Shortly describe the contribution of the project to each KPIs (from column "Current number of KPI")

Success of this project is measured by KPI2. However, as it is the development stage it is now measured as KPI2a.

CURRENT FINANCES

Estimated total cost	2,92 M NOK
Total expenditure by 2015-12-31	693 426 SEK
Planned expenditure by 2015-12-31	693 426 SEK

B5. ICT for data collection, safety and communication

Project Title	IT-system development for individualization of cervical cancer screening and implementation of new strategies to reach non-attenders
Project Leader	Sven Törnberg
Institution(s)	Regional cancer centre, Stockholm, Kreftregisteret, Karolinska Institutet
Date started	2015-01-01
Estimated completion date	2017-01-01

LIST OF PERSONNEL EMPLOYED BY THE PROJECT

Name	Position	Academic Degree	Gender
Agneta Carlsten Thor	Midwife		Female

CURRENT STATUS

Report on progress

The ability of screening programs to handle risk-stratified screening requires correct and functioning ICT tools, which are currently at least in part lacking. In a NIASC project, led by Ass. Professor Sven Törnberg, to see in real-life how we can adapt screening invitations and management for different groups of women we aim to send different letters to different women residing in the greater Stockholm area. Currently, women who have not attended screening are sent an annual reminder invitation. We want to examine whether alternative strategies (self-sampling or contact with a midwife) may work as well, or better, than the current strategy of just sending reminder invitations to non-attenders. The aim of this randomized health services study is therefore both to test the feasibility of providing alternative strategies instead of the customary reminder invitations to women in Stockholm (so-called feasibility study, primary aim), and whether such strategies will increase overall participation in screening (secondary aim). Offering self-sampling kits for human papillomavirus (HPV) testing to women who have not responded to screening invitations in the organized program, might lead to greater participation without substantial increases in the overall health care costs. A new IT software for ordering a self-sampling kit has been developed at the NIASC coordinating office, Karolinska Institutet. This software, in conjunction with the screening database at Regional Cancer Screening Center of Stockholm County, is used to identify and invite women who are long-term non-participants in cervical screening. Women who have not participated for ten years, despite annual reminder invitations, were identified and randomized to one of four strategies. First batch of invitation letters were sent to long term non-attenders on 14-03-2016. They have possibility to order self-sampling tests kits through the web page http://hemtest.hpvcenter.se.

PROJECT PERFORMANCE

Estimate project according NIASC Key Performance Indicators (KPIs)

Name	**Description**	Current number of KPI	Estimated number of KPI by the end of the project
KPI 1	***Risk stratification strategies using NIASC eScience tools actually implemented in national cancer screening program(s) for risk stratified screening.***	0	1
KPI 1a	*as above but in development or pilot implementation stage*	1	-
KPI 2	***Number of Open Source eScience tools developed***	1	1
KPI 2a	*as above but in development stage*	0	-
KPI 3	***Number of research datasets where open access is promoted using NIASC eScience tools***	0	0
KPI 3a	*as above but in development stage*	0	0
KPI 4	***Number of joint international projects***	0	0
KPI 5	***Academic Outputs*** *(Number of Papers published; Number of Workshops and Courses organized; Impact factor of the consortium)*	0	2

Shortly describe the contribution of the project to each KPIs (from column "Current number of KPI")

Currently, this project is contributing to KPI 1a as the project itself is in the trial phase and KPI 2, as open source web application for ordering self sampling kits is already developed and is currently in use.

CURRENT FINANCES

Estimated total cost	1 500 000 NOK
Total expenditure by 2015-12-31	186 137 NOK
Planned expenditure by 2015-12-31	186 137 NOK

B6. SOIGNonS: SOcIetal Games to Nudge people intO attending cervical caNcer Screening

Project Title	*SOIGNonS: SOcIetal Games to Nudge people intO attending cervical caNcer Screening*
Project Leader	Mari Nygård
Institution(s)	Cancer Registry of Norway, Icelandic Cancer Registry

Date started	15-04-2015
Estimated completion date	15-04-2017

LIST OF PERSONNEL EMPLOYED BY THE PROJECT

Name	Position	Academic Degree	Gender
Tomás Ruiz-López	Postdoctoral Fellow	Ph.D. Information and Communication Technologies	Male

CURRENT STATUS

Cervical cancer is a highly preventable cancer but only if women choose to be regularly screened. In virtually all countries, including Norway and Iceland, with an effective cervical cancer screening program, there is a small segment (~20%) of the population that do not participate in cervical cancer screening and consequently contribute significantly to the national burden of cervical cancer. SOIGNONS aims virally communicate health information concerning cervical cancer via gamification in mobile games.

In 2015 we established the SOIGNON`s working group at the Research Department at the Cancer Registry of Norway, and commenced several activities. The working group consists of postdoctoral fellow, computer scientists, screening experts in Norway, Iceland and US, research assistant. The Norwegian Cancer Society was contacted to leverage the app distribution at the time of the launch.

Shortly, game named “Fight HPV” with a purpose to present information on cervical cancer screening as thought-provoking puzzles to a large number of smart phone users, was designed. Characters, such as epithelial cells, high and low risk HPV, vaccines, prevention methods and screening were introduced in the game (figure 1). The player is presented, in an attractive way, pieces of information about the normal cervix, human papillomavirus (HPV), how to avoid infection with the virus, concepts of screening and vaccination in preventing HPV and cervical cancer during the course of the game. The information and the game experience complement each other during the learning process for the player. At the end of the play, it is possible to share results with acquaintances in order to spread engagement. The game will be published in the first half of 2016 tentatively, aiming to have media coverage to help us spread its availability. An article about the overview of the game was published in the Special Issue of Cancer in Norway 2014, in December 2015.

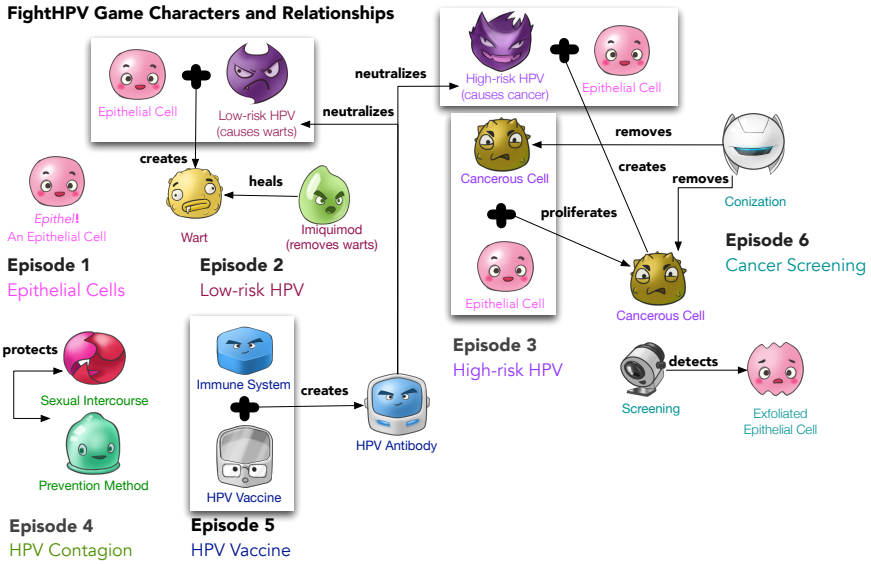


Figure 1. FightHPV Characters, Episodes and their Relationships

Project activities in 2015:

- **Game modelling and development**: we have designed a puzzle-based game with the main concepts regarding the interaction between the human papillomavirus and cervical disease. We introduce characters, such as epithelial cells, high and low risk HPV, vaccines, prevention methods and screening. This design has been implemented as a mobile game for Android and iOS devices (mobiles and tablets), consisting of 60 levels of increasing difficulty. The player is presented, in an attractive way, pieces of information about the previously mentioned concepts. The player has to solve a puzzle involving characters representing cells and viruses, which are driven by rules resembling their interactions in real life. In this way, the information and the game experience complement each other during the learning process for the player. At the end of the play, it is possible to share her results with acquaintances in order to spread engagement and get more persons to the game. We intend to leverage such coverage to let the general public know about the game, download it and play it.
- **Beta testing**: The closed beta testing for Android version was performed in nov-dec, in order to get early feedback by multiple users before releasing the game to the general public. The feedback that we obtain served to improve the app and its content, and fix any possible errors that may appear on certain devices. Users filled out a questionnaire after using the app that will help us conduct such tasks. Beta-testing of the iOS version in currently ongoing.
- **Ethical committee approval**: in parallel, we have worked on the research protocol which allows to evaluate the real-life impact of the game to the cervical cancer screening attendance. Our hypothesis is to raise the awareness about HPV and Cervical Cancer through the mobile game which will drive more people to Cervical Cancer Screening. We submitted an application to the Norwegian Regional Ethical Committee together with the protocol by 23rd September 2015 and received approval on 19th November 2015.
- **Data inspectorate approval**: in order to be able to apply the research protocol that we have designed, we need to access some participants' data. To be able to do so, we have submitted the necessary request to the Norwegian Data inspectorate by 24th November, 2015, which is still pending.
- **Article in Special Issue Cancer in Norway**: we published an article describing an overview of the game that will be published with the Cancer in Norway 2014, in December 2015.

Spinn-off project: Piloting "Fight-HPV" in Zambian adolescents. Zambia is a country with high Cervical cancer burden and where adolescents are in vulnerable situation in terms of the health

problems as compared to the general public. We hypnotised that game with serious content can be an appealing tool to introduce adolescent's reproductive health related concepts. In collaboration with the social scientists we developed a study protocol entitled "Human papillomavirus vaccination as platform for adolescent sexual and reproductive healthcare delivery in Zambia" and presented it to the Ethical Committee in 2015.

- **Ethical**

Significant changes to plan that have occurred (if applicable)

The original expected date to the first release of the mobile apps was December 2015. However, given that we didn't receive the approval from the Norwegian Data Inspectorate to conduct our research plan on time, we decided to postpone the release date until we get their approval.

Updated plan is provided as attachment. The project can be divided in three parts: i) app development; ii) collection storage of informed consent; iii) evaluation of the app effect on screening attendance. (figure 2)

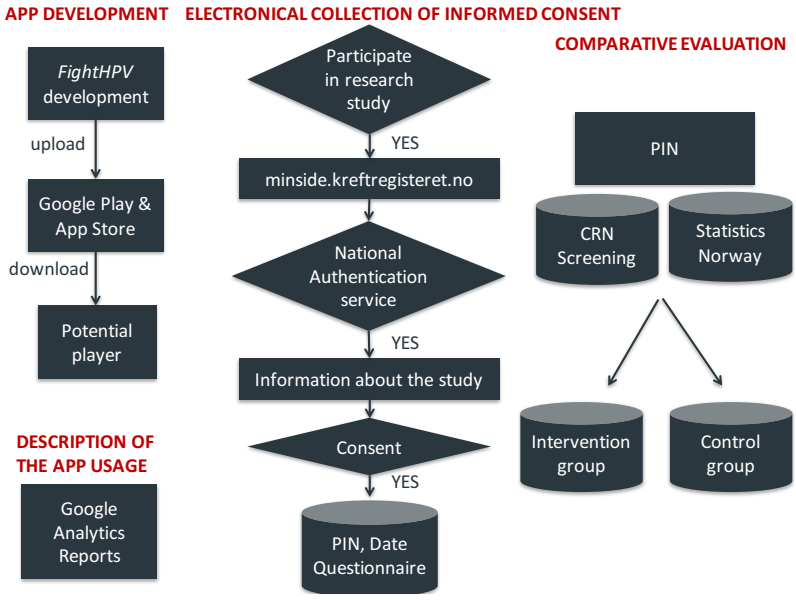


Figure 2. Three stages of the project.

PROJECT PERFORMANCE

Estimate project according NIASC Key Performance Indicators (KPIs)

Name	**Description**	Current number of KPI	Estimated number of KPI by the end of the project
KPI1	***Risk stratification strategies using NIASC eScience tools actually implemented in national cancer***	0	0

	screening program(s) for risk stratified screening.		
KPI1a	*as above but in development or pilot implementation stage*	0	-
KPI2	***Number of Open Source eScience tools developed***	1	1
KPI2a	*as above but in development stage*	0	-
KPI3	***Number of research datasets where open access is promoted using NIASC eScience tools***	0	0
KPI3a	*as above but in development stage*	0	-
KPI4	***Number of joint international projects***	0	0
KPI5	***Academic Outputs*** *(Number of Papers published; Number of Workshops and Courses organized; Impact factor of the consortium)*	1	2

Shortly describe the contribution of the project to each KPIs (from column "Current number of KPI")

Thus, success of this project is measure by KPI2, as it is already developed and available, as well as will be available as an open source app.

CURRENT FINANCES

Estimated total cost	1 509 000 NOK
Total expenditure by 2015-12-31	777 718 NOK
Planned expenditure by 2015-12-31	777 718 (23 037*) NOK

*Without salary expenses

C. Pilot projects testing eScience tools

Project Title	Data and text mining of cancer symptoms and comorbidities in electronic patient records in the Nordic languages, MINECAN
Project Leader	Hercules Dalianis
Institution(s)	Stockholm University, Technical University of Denmark, Cancer Registry of Norway
Date started	1 January 2015
Estimated completion date	31 December 2018

LIST OF PERSONNEL EMPLOYED BY THE PROJECT

Name	**Position**	**Academic Degree**	**Gender**
Rebecka Weegar	PhD Student	MSc	Female

CURRENT STATUS

Report on progress

A NIASC PhD-project led by Professors Soren Brunak and Hercules Dalianis, a Danish-Swedish collaboration, is using text mining analytic techniques to investigate potential additional predictive factors that could be used in cancer screening. Over a million patient records are mined for previously unknown symptoms and/or laboratory values that could/should be useful for risk stratification in screening programs. In January 1, 2015, a PhD position was filled. Ethical permission for the project was obtained to access the Karolinska University Hospital's electronic patient records. The health records were already stored in the HEALTH BANK - Swedish Health Record Research Bank at Stockholm University. A preliminary report is entitled "Finding Cervical Cancer Symptoms in Swedish Clinical Text using a Machine Learning Approach and NegEx". The study combined and evaluated two existing tools for text mining and gave insight into the needs for future work in this domain. Two rounds of manual annotation of existing medical records have been performed, and guidelines have been developed for the annotation. Immediate future plans are to include the new annotations to improve the performance of existing tools in the cervical cancer domain. Also, it is planned to apply for access to primary care data, which has been identified as potentially very valuable for this project. Another part of the project, mining Norwegian pathology reports, resulted in one paper entitled "Creating a rule based system for text mining of Norwegian breast cancer pathology reports". This paper was a feasibility study, where the possibility of a well-functioning system for information extraction was showed. The existing solution is rule-based and our future plan is to further develop this system and to include machine learning techniques and more data.

PROJECT PERFORMANCE

Estimate project according NIASC Key Performanc Indicators (KPIs)

Name	**Description**	Current number of KPI	Estimated number of KPI by the end of the project
KPI1	***Risk stratification strategies using NIASC eScience tools actually implemented in national cancer screening program(s) for risk stratified screening.***	0	0
KPI1a	*as above but in development or pilot implementation stage*	0	-
KPI2	***Number of Open Source eScience tools developed***	0	1
KPI2a	*as above but in development stage*	1	-
KPI3	***Number of research datasets where open access is promoted using NIASC eScience tools***	0	0
KPI3a	*as above but in development stage*	0	-
KPI4	***Number joint international projects***	1	1
KPI5	***Academic Outputs*** *(Number of Papers published; Number of Workshops and Courses organized; Impact factor of the consortium)*	5	10

Shortly describe the contribution of the project to each KPIs (from column "Current number of KPI")

Results of this project will be evaluated by NIASC KPI2 and has already contributed to KPI5 with 5 publications. As this project involves collaboration of 3 Nordic countries (Denmark, Norway and Sweden) it also contributes to KPI 4

CURRENT FINANCES

Estimated total cost	2 300 000 SEK
Total expenditure by 2015-12-31	1 031 899 SEK
Planned expenditure by 2015-12-31	1 031 899 SEK