*Journal of the American Medical Informatics Association*, 24(5), 2017, 950–957
doi: 10.1093/jamia/ocx038
Advance Access Publication Date: 21 April 2017
Research and Applications

Research and Applications

# E-Science technologies in a workflow for personalized medicine using cancer screening as a case study

**Ola Spjuth,[1,2] Andreas Karlsson,[1] Mark Clements,[1] Keith Humphreys,[1] Emma Ivansson,[1] Jim Dowling,[3] Martin Eklund,[1] Alexandra Jauhiainen,[1,4] Kamila Czene,[1] Henrik Grönberg,[1] Pär Sparén,[1] Fredrik Wiklund,[1] Abbas Cheddad,[1,5] Þorgerður Pálsdóttir,[1,6] Mattias Rantalainen,[1] Linda Abrahamsson,[1] Erwin Laure,[7] Jan-Eric Litton,[1,8] and Juni Palmgren[1,9]**

[1]Department of Medical Epidemiology and Biostatistics and Swedish e-Science Research Centre, Karolinska Institutet, Stockholm, Sweden, [2]Department of Pharmaceutical Biosciences and Science for Life Laboratory, Uppsala Universitet, Uppsala, Sweden, [3]School of Information and Communication Technology, KTH Royal Institute of Technology, Stockholm, Sweden, [4]Early Clinical Biometrics, AstraZeneca AB R&D, Gothenburg, Sweden, [5]Department of Computer Science and Engineering, Blekinge Institute of Technology, Karlskrona, Sweden, [6]Nordic Information for Action e-Science Center, Stockholm, Sweden, [7]School of Computer Science and Communication, KTH Royal Institute of Technology, Stockholm, Sweden, [8]Biobanking and Biomolecular Resources Research Infrastructure–European Research Infrastructure Consortium, Graz, Austria, and [9]Institute for Molecular Medicine Finland, Helsinki University, Helsinki, Finland

Corresponding Author: Ola Spjuth, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Box 281, SE-171 77 Stockholm, Sweden. E-mail: ola.spjuth@farmbio.uu.se. Phone: +46-70-425 06 28

Received 17 October 2016; Revised 13 February 2017; Accepted 17 March 2017

## ABSTRACT

**Objective:** We provide an e-Science perspective on the workflow from risk factor discovery and classification of disease to evaluation of personalized intervention programs. As case studies, we use personalized prostate and breast cancer screenings.
**Materials and Methods:** We describe an e-Science initiative in Sweden, e-Science for Cancer Prevention and Control (eCPC), which supports biomarker discovery and offers decision support for personalized intervention strategies. The generic eCPC contribution is a workflow with 4 nodes applied iteratively, and the concept of e-Science signifies systematic use of tools from the mathematical, statistical, data, and computer sciences.
**Results:** The eCPC workflow is illustrated through 2 case studies. For prostate cancer, an in-house personalized screening tool, the Stockholm-3 model (S3M), is presented as an alternative to prostate-specific antigen testing alone. S3M is evaluated in a trial setting and plans for rollout in the population are discussed. For breast cancer, new biomarkers based on breast density and molecular profiles are developed and the US multicenter Women Informed to Screen Depending on Measures (WISDOM) trial is referred to for evaluation. While current eCPC data management uses a traditional data warehouse model, we discuss eCPC-developed features of a coherent data integration platform.
**Discussion and Conclusion:** E-Science tools are a key part of an evidence-based process for personalized medicine. This paper provides a structured workflow from data and models to evaluation of new personalized intervention strategies. The importance of multidisciplinary collaboration is emphasized. Importantly, the generic concepts of the suggested eCPC workflow are transferrable to other disease domains, although each disease will require tailored solutions.

**Key words**: e-Science, cancer, personalized screening, data integration, modeling, simulation

## INTRODUCTION

The sequencing of the human genome opened avenues for new genomic analysis of diseases, and high-throughput technologies such as massively parallel sequencing and array-based technologies have resulted in new insights and new biomarkers. Despite these findings, the results have not yet propagated into new screening programs on a wide scale.

Biomarkers provide means of stratifying patients by disease risk, probability of therapy response, or prognosis. Genome-wide association studies are commonly used to detect genetic risk factors predisposing to complex diseases, including various cancers; however, clinical use when it comes to predicting individual risk is so far limited.[1] Large collaborative efforts such as the Cancer Genome Atlas and the International Cancer Genome Consortium aim to profile most major cancer forms using several molecular technologies and make data available to the research community. These studies contribute to establishing somatic mutation patterns, molecular subtypes, and pan-cancer molecular signatures.[2] In the clinical setting to date, the main patient benefit of cancer genomics has been to provide means for identifying patients harboring mutations in oncogenes such as EGFR and BRAF,[3,4] for which targeted therapies exist. A unique opportunity exists to develop and validate novel biomarkers for individual risk prediction, therapy response, and prognosis based on integration of multiple molecular data types and from multiple sources. However, new and refined bioinformatics, statistical methodologies, and other e-Science tools will be required to achieve these objectives.

The focus of this paper is an e-Science perspective on personalized prevention and treatment programs, describing and exemplifying the workflow from data and modeling to evaluation and clinical implementation. Substantial resources have been put into e-Science activities in many countries, including the United Kingdom[5] and the United States. Examples of e-Science approaches in the medical field include the work of Saltz et al.[6] who describe how e-Science can contribute to translational medicine from a theoretical computer science perspective, and Marias et al.[7] who describe an e-Science approach to multiscale simulation for the development of cancer therapies.

Each cancer form and subtype has its own risk factor profile, cancer genomics, prognosis, and treatment modalities. Currently, cancer screening programs for early detection use a one-size-fits-all approach instead of a personalized protocol, which takes into account a person's biological characteristics, circumstances, and preferences. The US National Cancer Intervention Surveillance Modeling Network (CISNET)[8] has made major contributions to understanding the effect of mass screening and treatment on trends in mortality. The CISNET collaboration started in 2000 from a cancer surveillance perspective and involves a number of cancers. Strengths of CISNET include a comparative modeling approach, with population-level reconstructions of risk exposures (eg, smoking for lung cancer) and screening behaviors (eg, mammography for breast cancer, prostate-specific antigen [PSA] testing for prostate cancer). CISNET, however, does not have access to the longitudinal data and complete life histories needed to design, implement, and evaluate new personalized screening programs.

As a conceptual case study, we present the approach taken by the Swedish program e-Science for Cancer Prevention and Control (eCPC), a flagship of the Swedish e-Science Research Centre (www.e-science.se). eCPC was established in 2011 with the aim of coordinating methodologies for new cancer biomarker discovery and developing new screening tests and programs using e-Science methods, ie, methods from mathematical, statistical, data, and computational sciences.

## THE ECPC WORKFLOW

For prostate cancer, no organized screening program exists today, and the challenge is to balance the benefit from tests for early detection (such as the PSA test) with harmful effects of screening (high rates of negative biopsies and overdiagnosis). For breast cancer, organized mammographic screening is implemented in most of the developed world.

When attempting to initiate or improve programs for early detection of disease, the starting point must be the observable biology, ie, the current understanding of the disease progression. This is referred to as the natural history of the disease, which typically involves parameters that are either not observable, such as the initial onset of disease, or are only partially observed at specific time points, such as the stage of disease at the time of diagnosis. Methodology adequate for latent or partially observable entities is useful in this context.

Based on the description of the natural history as it is known and on observed data, the aim of the personalized screening program in eCPC is to identify risk factors and evaluate their ability to discriminate high-risk from low-risk individuals in the population. An individual's risk level determines the screening intervention he or she receives. New predictors and new models are evaluated using relevant data sources, both in a controlled trial setting and on real-world populations using microsimulation and probabilistic calibration. When better discrimination is achieved and validated, then a new iteration of the workflow begins in order to continuously improve the screening program. This general methodology is captured in the eCPC workflow in Figure 1, which illustrates the process from data and modeling to evaluation of new intervention programs. This workflow requires tailored solutions for each disease domain. The starting point is the observable biology of the disease (node 3 of the figure). This guides the choice of data sources and the first-stage analytic approach.

## PERSONALIZED CANCER SCREENING

Recently, new validated predictors such as cancer susceptibility genetic markers and medical imaging have led to improvement in the discrimination accuracy of prediction models and the potential for new cancer screening programs.[9–12] We illustrate e-Science technology in the eCPC workflow in Figure 1 through 2 examples: personalized screening for prostate cancer and breast cancer.

### Prostate cancer

Prostate cancer is the most common cancer in Swedish men, with 10 000 new diagnoses and 2500 deaths annually. For prostate cancer, there are high levels of opportunistic screening, with no organized screening program. Current testing using the PSA test may be doing more harm than good. The major challenge is to predict which men will have more aggressive prostate cancer, and thus reduce overdiagnosis and overtreatment.

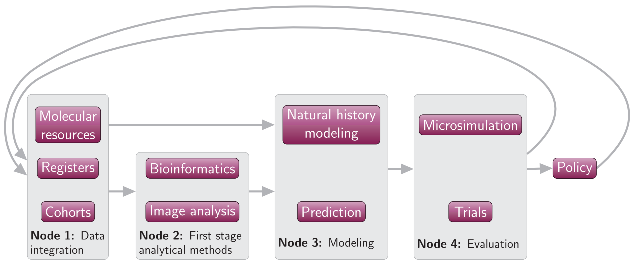


**Figure 1.** eCPC workflow to illustrate the process from data and modeling to evaluation of new population programs via 4 nodes. Prediction and natural history models are applied to assess individual risk. Model parameters are estimated using molecular data, nationwide Swedish registers, and cohort data. Bioinformatics and image analysis allow for discovery of novel biomarkers and other predictors in order to improve risk discrimination. Microsimulation is used to plan trials and evaluate protocols for public policy shifts. The process is iterative.

## Node 3: Natural history of prostate cancer

There are few environmental factors associated with prostate cancer incidence; the main risk factors are older age, family history of prostate cancer, and ethnicity. Natural history models for prostate cancer may include (1) cancer onset, stratified by the Gleason score used in treatment decisions, (2) stage progression for preclinical cancer, and (3) survival after diagnosis. The current prostate cancer natural history models vary in their mathematical representations, reflecting both parsimony and uncertainty in modeling for partially observed states. As examples, a model may assume no change of Gleason score after onset (eg, the model developed by the Fred Hutchinson Cancer Research Center [FHCRC][13]), or may model for de-differentiation (eg, the Microsimulation Screening Analysis[14] [MISCAN] model in CISNET). In addition, PSA can be modeled explicitly and linked to disease progression (eg, FHCRC), or PSA values can be modeled indirectly using test probabilities (eg, MISCAN). We adapted the FHCRC natural history model for the Swedish setting using Swedish data sources.

## Node 1: Data sources

Local studies include data from the Cancer Risk Prediction Center (CRisP)[15] for several prostate cancer studies, including an observational cohort, Stockholm-2, and a large diagnostic trial comparing biomarker effectiveness at predicting prostate cancer incidence, Stockholm-3 (STHLM3).[12] Through linkages of health registries, we organized a population-based registry, the Stockholm PSA and Biopsy Registry.[16] With data from the 3 clinical chemistry laboratories that perform all PSA analyses in the Stockholm region, we identified all men who underwent at least 1 PSA test in Stockholm since 2003. From the 3 pathology departments in Stockholm, data were collected on histological examinations from prostate tissue samples from the same geographical area and period. The register is regularly linked to the National Swedish Patient Register, where data on surgical procedures and histories regarding other diagnoses were obtained. Clinical data, including tumor stage and Gleason score, were obtained through linkages to the Regional Prostate Cancer Register and the Swedish National Cancer Register. For follow-up between 2003 and 2015, the Stockholm PSA and Biopsy Registry comprised 1.8 million PSA tests for 448 000 men. Extensive work has been done to normalize the Stockholm PSA and Biopsy database, and to annotate and standardize its use. For example, in order to ensure reproducible analyses, we developed data-curating procedures (eg, to clean biopsy data from misclassified surgeries based on surgery coding) and defined standardized definitions of how patients' PSA tests are associated with the prostate biopsies they undergo and the prostate cancer diagnoses they get.

## Node 2: Biomarker panel for prostate cancer risk prediction

Using data described in Node 1, we constructed an algorithm, the STHLM3 model (S3M), for predicting a man's risk of having clinically significant prostate cancer (defined as Gleason score 7 or higher) using a combination of plasma protein biomarkers (PSA, free PSA, intact PSA, human kallikrein-2, microseminoprotein beta, and macrophage inhibitory cytokine-1), genetic markers, clinical variables (age, family history, previous prostate biopsy), and a prostate exam (digital rectal exam and prostate volume) as predictors. The plasma protein biomarkers used in STHLM3 were selected from a scientific literature search and 2 subsequent validation studies on biobanked plasma (n = 1200 in total). S3M was fit to the STHLM3 training cohort (n = 11 130) using logistic regression.

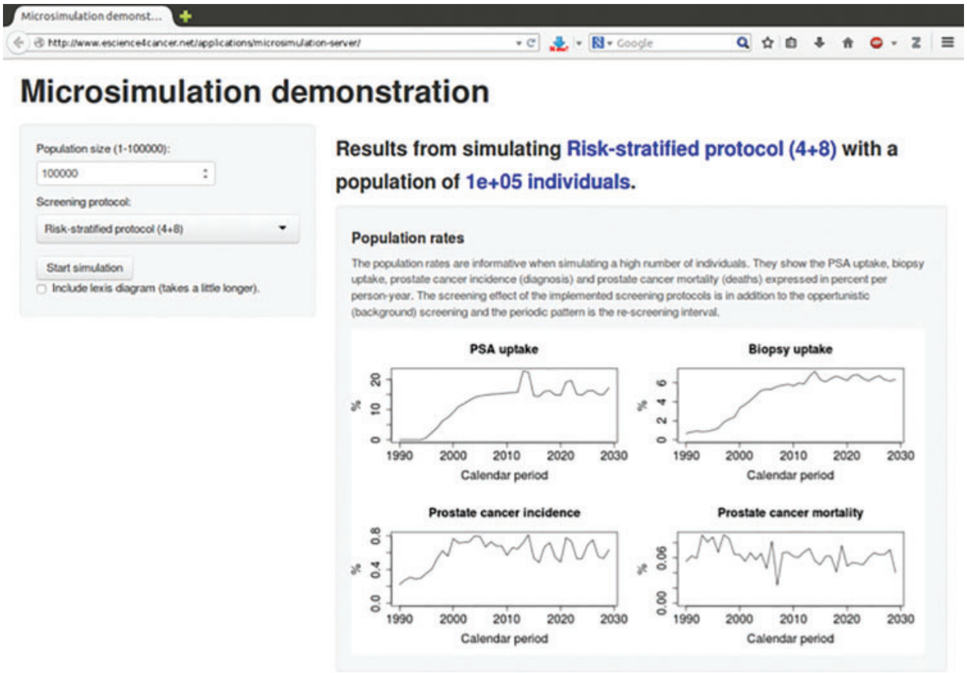
## Node 4: Controlled trial evaluation

S3M was tested in the validation cohort (no overlap with the training cohort) of the prospective and population-based STHLM3 diagnostic trial, which included 47 688 men ages 50–69 years.[12] The STHLM3 study followed a paired screen positive design, where 2 tests (PSA alone and S3M) were performed on each study participant. A paired design has considerably higher statistical power for a given number of study participants than a randomized design. The results of STHLM3 showed that S3M reduced the number of biopsies by 32% and avoided 44% of the negative biopsies compared to using PSA ≥ 3 ng/mL as a cutoff for biopsy recommendation, while maintaining the same sensitivity to diagnose Gleason score ≥7 as prostate cancer. S3M is now clinically available in Sweden and internationally and can be ordered instead of PSA as a clinical test for prostate cancer risk.

## Node 4: Microsimulation for evaluation in the population

Internationally, there are few open-source computational tools for population-based simulations of health care. The eCPC team developed an open-source microsimulation framework to simulate prostate cancer screening that closely links the R statistical language with a C++ microsimulation core used on high-performance computing clusters.[17] The simulation framework implements the prostate cancer simulation model originally developed by FHCRC. We extended the FHCRC model to incorporate S3M together with costs and quality-adjusted utilities for cost-effectiveness analysis. The source code is available on https://github.com/mclements/microsimulation and an online version (Figure 2) is available at http://www.escience4cancer.net/applications/microsimulation/.

To calibrate the microsimulation model, we explored the use of approximate Bayesian calibration.[18] For computationally expensive applications like the microsimulation, it is not feasible to evaluate the full likelihood motivating the use of approximate Bayesian calibration. Using the summary statistics and their likelihoods in combination has proved to quickly find suitable ranges for parameters in order to reproduce observable data. The Bayesian framework in itself is appealing for microsimulation model calibration, since prior information on suitable parameter ranges and distributions may be



Microsimulation demonst...
http://www.escience4cancer.net/applications/microsimulation-server/
Google

Microsimulation demonstration

Population size (1-100000): 100000

Screening protocol: Risk-stratified protocol (4+8)

□ Include lexis diagram (takes a little longer).

Results from simulating Risk-stratified protocol (4+8) with a population of 1e+05 individuals.

Population rates

The population rates are informative when simulating a high number of individuals. They show the PSA uptake, biopsy uptake, prostate cancer incidence (diagnosis) and prostate cancer mortality (deaths) expressed in percent per person-year. The screening effect of the implemented screening protocols is in addition to the opportunistic (background) screening and the periodic pattern is the re-screening interval.

**Figure 2.** Screen capture of the web-based microsimulation user interface for the prostate cancer model for a risk-stratified screening protocol, where men at low risk are rescreened every 8 years and men at medium risk are rescreened every 4 years.

available from previous studies or publications, and the posterior samples can be used for probabilistic sensitivity analysis.

### Clinical impact, outlook

Comparing observational and trial data, we found evidence that the S3M screening test would improve sensitivity and specificity in the population more than was estimated from the STHLM3 trial.[19] S3M began routine clinical use in Stockholm in 2016. The early use has been as a triage for the PSA test for intermediate PSA values. As suggested by a reviewer to,[19] we will be able to monitor the uptake and use of the S3M test in the population by extending the Stockholm PSA and Biopsy Register to include the S3M tests. A national prostate cancer testing register is being planned to extend the coverage across Sweden.

We have also begun the process of updating the S3M algorithm, allowing for revised and new biomarkers where we incorporate missing data methods and methods to include covariates only available for those who screened positive (eg, for those referred to biopsy).

## BREAST CANCER

Breast cancer is the most common cancer type among women in Sweden and accounts for approximately 15% of all female cancer deaths. For breast cancer, mammographic screening was introduced gradually across Sweden and national coverage was achieved in 1997. Women are offered screening between the ages of 40 and 74, 40 and 69, 46 and 69, 50 and 69, or 50 and 74, at intervals in the range of 18–24 months, depending on county of residence. Participation is in the range of 70–90%, again depending on county of residence. However, women continue to present clinically with breast cancer in the time intervals between screenings and outside of the screening program.

### Node 3: Natural history of breast cancer

Understanding screening sensitivity and tumor progression is an important part of the eCPC program for evaluating personalized screening programs for breast cancer. We are developing a framework for modeling tumor growth as a continuous biological process and have, for example, used this to estimate mammographic screening sensitivity as a function of tumor characteristics and mammographic density and to study determinants of tumor growth rates.[20,21]

### Node 1: Data sources

eCPC works with data from CRisP[15] also for breast cancer studies, including the Karolinska Mammography (KARMA) project for risk prediction of breast cancer,[22] a prospective cohort study comprising >70 000 women attending mammography screening or clinical mammography in Sweden. Participants have responded to web-based questionnaires, donated blood, and given permission for storage of full-field digital mammograms. eCPC also works with the Linné-Bröst 1 (LIBRO-1) study,[21] a large cohort of breast cancer patients in the Stockholm-Gotland region who were diagnosed with breast cancer between 2001 and 2008, and CAHRES,[17] a population-based case-control study of post-menopausal breast cancer, with 6000 cases and controls, of which half have been included in molecular studies.

### Node 2: Image-based markers for risk

Mammographic percent density has been confirmed to be one of the strongest risk factors for breast cancer in women, and solutions already exist to measure it in the clinical setting[23] and include it in risk-prediction models. There is, however, an opportunity to improve risk prediction by extracting further attributes from digital mammograms. This is being explored by a number of research programs, such as the WISDOM trial[24] and the European Commission

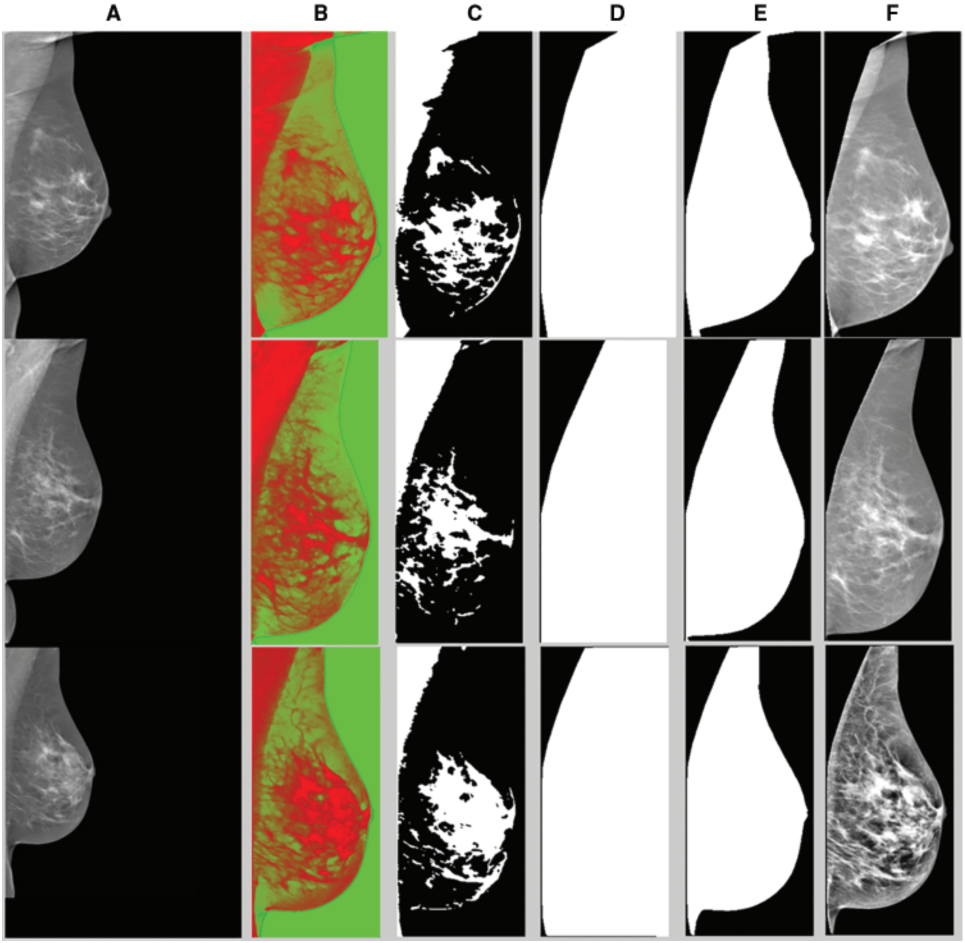


**Figure 3.** Automated mammography breast segmentation and feature extraction for breast cancer research. The figure shows the output of our preprocessing of mammograms: (**A**) original full-field digital mammograms, (**B**) pseudo-color generation after applying the horizontal and vertical cropping, (**C**) positive signal in the Q component in the NTSC color space, detecting the reddish area, (**D**) convex hull of the negative (c), (**E**) final extracted breast mask, and (**F**) breast region after applying the contrast limited adaptive histogram equalization. Note that to get the dense tissue region, one could perform a logical AND operation of the input images (**C** and **D**).

FP7-funded Assure.[25] A mammogram-based measure of breast tissue stiffness and local texture patterns in prediagnostic mammograms have, for example, recently been shown to be associated with breast cancer risk.[26],[27]

eCPC has recently developed a novel approach to enhance area-density measures in digitized film and digital mammograms[28],[29] and described algorithms (based on automated segmentation and machine learning methods; Figure 3) for measuring area and volumetric density in processed full-field digital mammograms[30] and spatial relations of density regions.[31] Additionally, and in a more recent study, we identified 2 novel mammographic features in screening mammograms of the normal breast that appear to differentiate between future-interval cancers and screen-detected cancers. The 2 image-based features are "skewness of the intensity gradient" and "eccentricity," which are basically shape descriptors. The former feature was found to be associated with tumor size at screen detection, while the latter was found to be associated with tumor size at interval detection.[32]

## Node 2: Genomic profiling to identify high risk

eCPC aims to identify molecular features that predict interval breast cancers, ie, tumors that become detectable in the interval between mammography screenings. Interval cancers represent a failure of the current screening system and are thought to be aggressive tumors with a high growth rate. We evaluated the genomic profiles of 60 interval and 113 screening-detected cancers through targeted deep sequencing of cancer-related genes, low-pass whole-genome sequencing, and RNA sequencing, in order to better understand somatic differences between the 2 groups.[33] Comparisons included routine tumor characteristics currently used in the clinic as well as the spectrum of breast cancer–associated point mutations, copy number alterations, and differential gene expression. The genomic profiling indicates that the genes TP53, PPP1R3A, and KMT2B are more frequently mutated in interval cancers than in screen-detected cancers. Adjustment for the PAM50 subtype indicates that the molecular differences by interval cancer status are largely explained by PAM50 subtypes, confirming that interval cancers are associated with more aggressive tumor characteristics and the identified genes are targets for identification of high-risk individuals.

## Node 4: The WISDOM trial

Traditionally, age has been the sole risk factor used to initiate screening. A personalized approach to screening could be an alternative to the current "one-size-fits-all" guideline-based approach in the United States and most of the rest of the developed world. Besides work on identifying new risk factors as described under eCPC Node 2, eCPC members are involved in the WISDOM study of risk, a multicenter trial comparing risk-based screening to annual screening in 100 000 women ages 40–74, initiated at the Athena Breast Health Network (in California and the Midwest at Sanford Health).[24] The trial will determine if screening based on personalized risk is as safe, results in less morbidity, is preferred by women, and will facilitate prevention for those most likely to benefit and adapt as we learn who is at risk for what kind of cancer. In the planning of WISDOM, a microsimulation model was developed to study, a priori, the likely outcomes of the trial, to determine the best approach for analyzing the trial, and to estimate study power and the likely cost-effectiveness of the intervention.

### Clinical impact, outlook

Accurate risk prediction is a necessary prerequisite for effective personalized breast cancer screening. If successful, the WISDOM study could provide a paradigm shift, with a decrease in the mortality of breast cancer and reduced unintended consequences of screening, similar to those for prostate cancer (unnecessary biopsies, overdiagnosis, and overtreatment).

## E-SCIENCE METHODS FOR DATA INTEGRATION

Here we describe some generic e-Science concepts and methods for data integration that are relevant in the context of the eCPC workflow. While eCPC has used traditional data management tools so far, the methods in this section are developed within eCPC aiming to adopt a more mature future data science approach. This section refers to Node 1 of the eCPC workflow in Figure 1.

Medical research commonly involves integrating data from various sources, such as biobanks, health registers, medical records, and local as well as public repositories. Data integration is the process by which data from different sources are linked and made available in a form such that information or conclusions can be deduced from the combined data.[34] This commonly requires that a metadata model exists, allowing data to be harmonized or transformed and converted into an interoperable format. Data integration can be achieved on different levels: as aggregated summary statistics or on the microdata (patient or sample) level. Integrating data is not trivial, due to issues of semantics, and also for political and regulatory reasons. The sensitive nature of information in medical research also calls for solutions using adequate anonymization and encryption methods to protect the confidentiality of individuals.[35]

### Integrating samples across biobanks and clinical health registers

We developed and applied the stereo-array isotope labeling (SAIL) method[36] for linking data across biobanks and health registers. SAIL operates on the metadata level, specifically availability data, and contains information on whether data for individual samples exist without disclosing the data value per se. This approach avoids many privacy issues but allows for linking data on the sample level to carry out power calculations or feasibility studies. The method is particularly useful when designing large-scale studies on a specific topic as well as in raising awareness among researchers about the availability and content of data sources, making the data easy to locate, interpret, and incorporate into a research project. We applied the SAIL method to biobank data in the European Network for Genetic and Genomic Epidemiology consortium[37] to local biobanks at Karolinska Institutet and to prostate cancer health registry data at the Karolinska University Hospital.[36,38] In order to reduce the chance of possible reidentification of individuals using prior information, we applied statistical disclosure control methods that have been successfully applied in several bioscience projects, such as Jerboa[39] and DataSHIELD.[40] Public instances demonstrating the method are available from http://www.escience4cancer.net/applications/data-availability/. The source code for SAIL is available from the SIMBioMS website, https://www.simbioms.org. However, the importance lies not in the actual software but rather in the developed general method of achieving interoperable data that can be applied to other similar cases and using other software tools.

### Secure integration of sensitive data in distributed environments

A common solution in medical data integration is to pseudonymize personal IDs (PIDs), which is associated with the risk of deidentification if prior information is available that can be linked via nonsensitive variables. We developed a method that replaces the PID with an anonymous ID, which is then subjected to double encryption using a Study key as well as a Master key. This ensures that data can be integrated on an anonymous ID in a system containing the Master key, but also ensures that if the system were compromised, the data could not be traced back to the original PID. The data are subsequently made available in an isolated secure instance of a data availability system[41] and can serve as a model for reducing the risk of deidentification when publishing availability data. The source code for the implementation is available at https://github.com/gholamiali/ecpc.

### Standardization

Developing shared vocabularies, minimum-information standards,[42] and data exchange formats is a key aspect of data integration carried out in various domain-specific settings as well as in international consortia, eg, the Biobanking and Biomolecular Resources Research Infrastructure–European Research Infrastructure Consortium,[43] the Public Population Project in Genomics and Society,[44] and BioMedBridges.[45] Vocabularies and standards emerge that gain acceptance in the medical research community, eg, Health Level 7, Clinical Data Interchange Standards Consortium, and Systematized Nomenclature of Medicine–Clinical Terms.

## DISCUSSION

This paper exemplifies e-Science tools useful in the process of personalized medicine. A generic workflow is defined starting from data and models and carries over to evaluation of programs in controlled and real-world settings. To illustrate e-Science methodology, we use personalized prostate and breast cancer screening. Importantly, the workflow in Figure 1 is not limited to cancer, but is adaptable to other disease domains.

Natural history cancer modeling will continue to be central to the validity of most simulation models for cancer screening and cancer treatment. The comparative modeling approach of CISNET brought consolidation to divergent research on cancer screening.[46] We now see a more mature field where natural history models are used to answer more refined questions about cost-effectiveness,[47] risk stratification, and evaluation of novel biomarkers.[48]

We expect to see improvements in the accuracy of *prediction models* through better and more affordable biomarkers. We also expect to see a trend toward prediction models for biological *subtypes* of cancer and changes in the response variable from pathologically determined phenotypes to phenotypes determined by genome sequencing. In the era of personalized medicine, large-scale prospective trials of personalized approaches to screening and treatment will likely become more common, leading to broader clinical adoption of prediction models.

With larger sample sizes and high-throughput technologies, both image-based and biomarker-based discovery have become data-intensive and inherently dependent on e-infrastructures and e-Science components. Similarly, in silico modeling and microsimulation are data-intensive and offer a flexible complement to conventional randomized intervention trials with morbidity or mortality outcomes. Although total mortality is an ultimate measure of intervention success, intermediate outcomes such as number of biopsies in the S3M diagnostic trial are important as intermediate evaluation tools. Moreover, the flexibility of the microsimulation framework allows for evaluation of cancer screening protocols based on cost-effectiveness comparisons and where the lifetime expected quality-adjusted utilities and costs can be calculated.[49]

Data sizes in, eg, eCPC are considerable and expected to increase manifold in the coming years with, eg, increased uptake of high-throughput technologies such as massively parallel sequencing. This, coupled with simulations of large populations, necessitates access to high-performance e-infrastructures (compute clusters or cloud resources) and requires that problems are properly parallelized. New emerging frameworks for Big Data promise to simplify construction of scalable software applications but are not easily accessible, and in most cases require considerable changes to existing code and libraries. Building a general modular system, such as the workflow in Figure 1, requires expertise in software engineering, high-performance and distributed computing, and collaboration with a multidisciplinary team of scientists.

A challenging and far-reaching implication of eCPC has been in cross-disciplinary development. A broad range of methods experts with backgrounds in mathematical, statistical, data, and computational sciences work closely together and interact on a daily basis with clinicians, epidemiologists, and molecular scientists. We must invest in developing the human competences necessary to realize these new approaches to doing science.[50,51] This is particularly true of the sciences that are traditionally less technical, which includes the biomedical sciences. Importantly, we also need to transform the way in which students in medicine view research.

## Acknowledgments

We thank the Swedish eScience Research Center community for numerous discussions, as well as Arnoldo Frigessi, University of Oslo, and Niklas Blomberg, Elixir UK, for excellent input as international eCPC Scientific Advisory Board members.

## COMPETING INTERESTS

AJ is an employee and shareholder of AstraZeneca.

## FUNDING

This work was supported by the Swedish e-Science Research Centre, the Swedish Research Council, the Nordic Information for Action eScience Center, the Swedish Cancerfonden (grant numbers CAN 2012/765, CAN 2014/472), Forte, and KID PhD funding from Karolinska Institutet.

## CONTRIBUTORSHIP

eCPC is a project where all authors contribute to multidisciplinary efforts, and in particular:

OS, JD, TP: data integration; EI, MR: biomarker discovery; KH, AC: image analysis; MC, AK, FW, ME, HG: prostate cancer applications; KH, LA, KC, AC: breast cancer applications; MC, AK: microsimulation; AJ: model calibration; JP, JEL, OS, MC, KH, PS, EL: strategic planning.

## REFERENCES

1. Manolio TA. Bringing genome-wide association findings into clinical use. *Nat Rev Genet*. 2013;14:549–58.
2. Chang K, Creighton CJ, Davis C, *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45:1113–20.
3. Siena S, Sartore-Bianchi A, Di Nicolantonio F, *et al*. Biomarkers predicting clinical outcome of epidermal growth factor receptor-targeted therapy in metastatic colorectal cancer. *J Natl Cancer Inst*. 2009;101:1308–24.
4. Jang S, Atkins MB. Which drug, and when, for patients with BRAF-mutant melanoma? *Lancet Oncol*. 2013;14:e60–69.
5. Research Councils United Kingdom. *RCUK Review of e-Science 2009*. 2009. https://www.epsrc.ac.uk/newsevents/pubs/rcuk-review-of-e-science-2009-building-a-uk-foundation-for-the-transformative-enhancement-of-research-and-innovation/. Accessed October 4, 2016.
6. Saltz J, Kurc T, Hastings S, *et al*. e-Science, caGrid, and translational biomedical research. *Computer (Long Beach, Calif)*. 2008;41:58–66.
7. Marias K, Dionysiou D, Sakkalis V, *et al*. Clinically driven design of multi-scale cancer models: the ContraCancrum project paradigm. *Interface Focus*. 2011;1:450–61.
8. National Cancer Institute–Cancer Intervention and Surveillance Modeling Network (CISNET). http://cisnet.cancer.gov. Accessed October 4, 2016.
9. Darabi H, Czene K, Zhao W, *et al*. Breast cancer risk prediction and individualised screening based on common genetic variation and breast density measurement. *Breast Cancer Res*. 2012;14:R25.
10. Vachon CM, Pankratz VS, Scott CG, *et al*. The contributions of breast density and common genetic variation to breast cancer risk. *J Natl Cancer Inst*. 2015;107(5): dju397.
11. Cuzick J, Brentnall AR, Segal C, *et al*. Impact of a panel of 88 single nucleotide polymorphisms on the risk of breast cancer in high-risk women: results from two randomized tamoxifen prevention trials. *J Clin Oncol*. 2017;35(7):743–50.
12. Grönberg H, Adolfsson J, Aly M, *et al*. Prostate cancer screening in men aged 50–69 years (STHLM3): a prospective population-based diagnostic study. *Lancet Oncol*. 2015;16:1667–76.
13. Gulati R, Gore JL, Etzioni R, *et al*. Comparative effectiveness of alternative prostate-specific antigen–based prostate cancer screening strategies. *Ann Intern Med*. 2013;158:145.
14. Heijnsdijk EAM, de Carvalho TM, Auvinen A, *et al*. Cost-effectiveness of prostate cancer screening: a simulation study based on ERSPC data. *J Natl Cancer Inst*. 2014;107:dju366–dju366.
15. The Cancer Risk Prediction Center (CrisP). http://ki.se/en/meb/crisp. Accessed October 4, 2016.
16. Nordström T, Aly M, Clements MS, *et al*. Prostate-specific antigen (PSA) testing is prevalent and increasing in Stockholm County, Sweden, despite no recommendations for PSA screening: results from a population-based study, 2003–2011. *Eur Urol*. 2013;63:419–25.
17. Karlsson A, Olofsson N, Laure E, *et al*. A parallel microsimulation package for modelling cancer screening policies. In: *2016 IEEE 12th International Conference on eScience*. Baltimore, MD: (e-Science); 2016, 323–30.
18. Rutter CM, Miglioretti DL, Savarino JE. Bayesian calibration of microsimulation models. *J Am Stat Assoc*. 2009;104:1338–50.
19. Eklund M, Nordström T, Aly M, *et al*. The Stockholm-3 (STHLM3) model can improve prostate cancer diagnostics in men aged 50–69 yr compared with current prostate cancer testing. *Eur Urol Focus*. 2016.

20. Abrahamsson L, Czene K, Hall P, *et al*. Breast cancer tumour growth modelling for studying the association of body size with tumour growth rate and symptomatic detection using case-control data. *Breast Cancer Res*. 2015;17:116.
21. Abrahamsson L, Humphreys K. A statistical model of breast cancer tumour growth with estimation of screening sensitivity as a function of mammographic density. *Stat Methods Med Res*. 2016;25:1620–37.
22. The KARMA Study. http://karmastudy.org. Accessed October 4, 2016.
23. Highnam R, Brady SM, Yaffe MJ, *et al*. *Robust Breast Composition Measurement – VolparaTM*. Springer: Berlin, Heidelberg; 2010:342–9.
24. Shieh Y, Eklund M, Madlensky L, *et al*. Breast cancer screening in the precision medicine era: risk-based screening in a population-based trial. *J Natl Cancer Inst*. 2017;109:djw290.
25. The Assure project. http://www.assure-project.eu/. Accessed October 4, 2016.
26. Boyd NF, Li Q, Melnichouk O, *et al*. Evidence that breast tissue stiffness is associated with risk of breast cancer. *PLoS One*. 2014;9(7):e100937.
27. Nielsen M, Vachon CM, Scott CG, *et al*. Mammographic texture resemblance generalizes as an independent risk factor for breast cancer. *Breast Cancer Res*. 2014;16:R37.
28. Cheddad A, Czene K, Shepherd JA, *et al*. Enhancement of mammographic density measures in breast cancer risk prediction. *Cancer Epidemiol Biomarkers Prev*. 2014;23:1314–23.
29. Cheddad A, Czene K, Hall P, *et al*. Pectoral muscle attenuation as a marker for breast cancer risk in full-field digital mammography. *Cancer Epidemiol Biomarkers Prev*. 2015;24:985–91.
30. Cheddad A, Czene K, Eriksson M, *et al*. Area and volumetric density estimation in processed full-field digital mammograms for risk assessment of breast cancer. *PLoS One*. 2014;9:e110690.
31. Ali MA, Garnier M, Humphreys K. Spatial relations of mammographic density regions and their association with breast cancer risk. *Procedia Comput Sci*. 2016;90:169–74.
32. Strand F, Humphreys K, Cheddad A, *et al*. Novel mammographic image features differentiate between interval and screen-detected breast cancer: a case-case study. *Breast Cancer Res*. 2016;18:100.
33. Li J, Ivansson E, Klevebring D, *et al*. Molecular differences between screen-detected and interval breast cancers are largely explained by PAM50 subtypes. *Clin Cancer Res*. 2016.
34. Solomon DJ, Henry RC, Hogan JG, *et al*. Evaluation and implementation of public health registries. *Public Health Rep*. 1991;106:142–50.
35. Reiter JP, Kinney SK. Commentary: sharing confidential data for research purposes. *Epidemiology*. 2011;22:632–35.
36. Spjuth O, Krestyaninova M, Hastings J, *et al*. Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research. *Eur J Hum Genet*. 2015;41:1–8.
37. The ENGAGE consortium. http://www.euengage.org. Accessed October 4, 2016.
38. Spjuth O, Heikkinen J, Litton J-E, *et al*. *Data Integration between Swedish National Clinical Health Registries and Biobanks Using an Availability System*. Springer International Publishing; 2014:32–40.
39. Avillach P, Coloma PM, Gini R, *et al*. Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. *J Am Med Inform Assoc*. 2013;20(1):184–92.
40. Wolfson M, Wallace SE, Masca N, *et al*. DataSHIELD: Resolving a conflict in contemporary bioscience — performing a pooled analysis of individual-level data without sharing the data. *Int J Epidemiol*. 2010;39:1372–82.
41. Gholami A, Laure E, Somogyi P, *et al*. Privacy-preservation for publishing sample availability data with personal identifiers. *J Med Bioeng*. 2015;4(2):117–25.
42. Taylor CF, Paton NW, Lilley KS, *et al*. The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol*. 2007;25:887–93.
43. BBMRI.eu. http://www.bbmri-eric.eu/. Accessed October 4, 2016.
44. P3G. http://p3g.org/. Accessed October 4, 2016.
45. BioMedBridges. http://www.biomedbridges.eu/. Accessed October 4, 2016.
46. Berry DA, Cronin KA, Plevritis SK, *et al*. Effect of screening and adjuvant therapy on mortality from breast cancer. *N Engl J Med*. 2005;353:1784–92.
47. Heijnsdijk EAM, de Carvalho TM, Auvinen A, *et al*. Cost-effectiveness of prostate cancer screening: a simulation study based on ERSPC data. *J Natl Cancer Inst*. 2015;107:366.
48. Birnbaum JK, Feng Z, Gulati R, *et al*. Projecting benefits and harms of novel cancer screening biomarkers: a study of PCA3 and prostate cancer. *Cancer Epidemiol Biomarkers Prev*. 2015;24:677–82.
49. Hunink MGM, Weinstein MC, Wittenberg E, *et al*. *Decision Making in Health and Medicine: Integrating Evidence and Values*, 2nd ed. Cambridge: Cambridge University Press; 2014.
50. Community cleverness required. *Nature*. 2008;455:1–1.
51. Stein LD. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nat Rev Genet*. 2008;9:678–88.