# CompBioMed Conference 2019.
# *New areas for innovative applications of computing to cancer*
# Towards personalised cancer prevention: The Digital Cancer Precision Prevention Initiative

Nygård, M.[1], Soper, B. [2], , Abdulla, G. [2], Sales, AP. [2], Ray, P. [2], Widemann, D. [2], Goncalves, A. [2], Campbell, S.[1] Nygård, J.F.[1],

[1]Cancer Registry of Norway, Oslo, Norway

[2]Lawrence Livermore National Laboratory, Livermore, CA

## 1. Background

We live in the information age where the flow of knowledge, including medical advice and innovations, quickly reaches each of us. However, existing recommendations for disease prevention, diagnostics and treatment are population-based, or based on highly selected randomized controlled trials, and only seldomly account for individual differences. For effective control of globally increasing morbidity and mortality due to cancer, the focus on early detection and intervention cannot be underestimated. The estimated spiraling costs of cancer treatment will challenge even the highest-income countries and underline the urgent need to develop preventive efforts.

Knowledge of biological disease mechanisms along with existing individual data from national population-based health registries, biobanks and surveys can be tailored for personally designed actions safely, efficiently and quickly.

Cervical cancer screening is an excellent model system for the development of personalised strategies for cancer prevention. It has a proven strong effect for decreasing cancer burden at the population level, and the Norwegian population-based screening program has produced large amounts of individual data that is accessible by centrally organized nationwide registries.

## 2. Objectives

The overarching aim of this paper is to develop methodology for improved screening outcome prediction: i) modelling of the multi-stage morphological changes leading to cervical cancer; ii) predicting future disease states; iii) predicting cervical cancer screening adherence.

## 3. Data

The Cancer Registry of Norway has run a national cervical cancer screening program since October 1991, collecting all screening and diagnostic results. Though screening guidelines exist (e.g. a cytology smear every three years from age 25 to 69), screening is at the discretion of the individual woman. As a result, the number of screening records and the time between screenings vary considerably between women. Three types of exams are used in the screening program: cytology, histology and molecular tests detecting presence of high-risk types of human papillomavirus (HPV) DNA.

Cervical cancer screening outcomes are dependent on features which are beyond the scope of medical exams. In particular, personal lifestyle can be predictive of women's compliance with

cervical screening guidelines and exposure to the main causal agent, sexually transmitted HPV. 21,563 women of ages between 18 and 45 years were randomly selected from the general female population in 2004 and 2011 and invited to fill out a questionnaire with information on education, marital status, smoking history, alcohol intake, sexual habits, contraceptive use, sexually transmitted diseases, and reproductive history. The questionnaire data were linked to the Norwegian Cervical Cancer Screening Program databases, obtaining for each person information on every cervical screening exam between 1992 and 2016.

## 4. Models

According to the natural history, cervical cancer is an infrequent end-stage of minor cellular abnormalities caused by an HPV infection. These abnormalities progress from minor changes, through more definitely premalignant changes, to localized invasive cancer. If left untreated this can lead to metastatic disease and ultimately death. Being able to detect cervical cancer in its early stages, or pre-cancers, followed by prompt, appropriate treatment is the key element which justifies nation-wide cancer screening programs.

A necessary step towards personalised screening programs is the development of reliable predictive models that take into account both disease dynamics and individual patient data.

We developed a continuous-time, time-inhomogeneous hidden Markov model reflecting cervical cancer carcinogenesis and the screening process. By leveraging 1.7 million individual's multivariate time-series with almost 11 million medical exams collected over a 25-year period, we estimated transmission parameters and predictive probabilities for a high-risk/cancer state at the next screening.

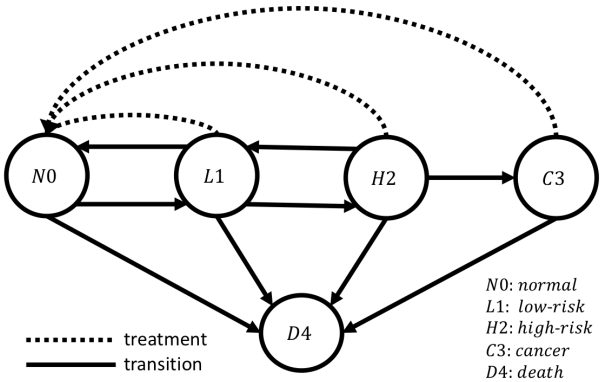


Figure 1 shows a continuous-time Markov model of cervical cancer development with consecutive states of normal, low-risk, high- risk, cancer and "death". The solid lines represent possible transitions in the Markov chain while dashed lines represent instantaneous resets due to treatment

A hierarchical Bayesian MTL approach, referred to as Bayesian Multitask with Structure Learning (BMSL) was used to combine life-style data collected through questionnaires to

predict screening adherence. This mode leverages commonalities across related tasks, i.e. life style events, with the aim of improving individual task performance. A key modeling choice in designing MTL models is the structure of the tasks' relatedness, which may not be known.

## 5. Results

Data engineering

To create a rich temporal dataset, health registry data was fused with survey data. As an outcome we had a single time series per patient with N time series where N is the number of patients. To protect the privacy, health data were manipulated, by collapsing test times within a specific window into a reference date and dates within a sample were shifted.
The impact of missing and erroneous data was assessed.

Modelling

We show that the HMM reflects the Norwegian screening program by comparing empirical survival curves for both registry data and data simulated from the proposed model. Calibration showed that using no historical data results in predictions very close to the population incidence rate and severely over and under estimated the true risk for patients that are at high risk. By utilizing individual screening histories and covariate data, the model shows potential for improving strategies for cancer screening programs by personalizing recommended screening intervals.

The BMSL model outperformed single task learning models in terms of predictive performance for screening adherence and performs at least as well as other MTL methods.

We continue to improve current models by including HPV status as covariate data, information on HPV vaccination status, incorporate lifestyle survey data as covariate data, develop hierarchical and nonparametric Bayesian models, utilize Recurrent Neural Networks (e.g. LSTM) to perform sequence prediction on patient test results with survey data as input.

## 6. Conclusions

We demonstrate the feasibility of developing personalised algorithms for cervical cancer prevention. The models can be generalized to include more detailed individual-level covariates as well as new types of screening exams. The ultimate goal is to develop models for predicting patient states to improve cancer screening outcomes and move towards a paradigm of personalised cancer prevention which complements existing population-based approaches.