UNIVERSITY
OF OSLO

Severin Elvatun

Towards more personalized cervical cancer prevention

Prototyping data-driven methods predicting cervical cancer development from cancer registry data

Thesis submitted for the degree of Philosophiae Doctor

Department of Informatics
Faculty of Mathematics and Natural Sciences

Cancer Registry of Norway

2023

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 2636

ISSN 1501-7710

Cover: UiO.
Print production: Graphic center, University of Oslo.

This PhD project was supported by the Decipher project, through the IKTPLUSS-program of the Research Council of Norway (grant number 300034).

To my ghostwriter

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of *Philosophiae Doctor* at the University of Oslo. The study has been approved by the Regional Committee for Medical and Healthcare Research Ethics (ref. 2015/1344). The work was carried out partially at my different home offices, the University of Oslo, the Cancer Registry of Norway (CRN) and SimulaMet. The thesis is a collection of three papers, presented in chronological order of writing. The common theme to them is the development and evaluation of data-driven methods predicting cervical cancer development from routinely collected exam history data, made available through the CRN.

Acknowledgements

Thinking back to my high-school days, the idea of venturing in any academic direction seemed far-fetched, although I am now very grateful for having had the opportunity granted by my co-supervisor Dr Valeriya Naumova and main supervisor Dr Jan F Nygård to pursue more matured interests. Profound gratitude is expressed for their support, inspiration and guidance.

Although almost two-thirds of this work were carried out at the home office, I got to know several wonderful people over the past three years. Without providing an exhaustive list, I would like to direct a special thanks to Dr Mari Nygård and Dr Markus Grasmair for their help and support.

Additionally, I am very thankful for the collaboration with Dr Evrim Acar, Dr Vinay C Gogineni, Mikal M Stapnes and Florian Becker. I appreciate that you all took the time to help, shared your thoughts and your ideas.

I am grateful for having been a part of the DeCipher project. Working in such an interdisciplinary environment has been educational and compelling. Thanks for the dinners and the expensive wine!

Moreover, my gratitude extends to my friends and colleagues at the Cancer Registry of Norway and SimulaMet. The social environment at work (and spending the afternoons at the climbing centers) was always much appreciated.

Last, but never least, I would like to thank my family for their unceasing support and patience in all my endeavours.

Severin Elvatun
Oslo, July 2023

List of Papers

Paper I

Langberg, GSRE et al. "Matrix factorization for the reconstruction of cervical cancer screening histories and prediction of future screening results". *Published in BMC Bioinformatics.* (2022), DOI: https://doi.org/10.1186/s12859-022-04949-8.

Paper II

Langberg, GSRE et al. "Towards a data-driven system for personalized cervical cancer risk stratification". *Published in Scientific Reports.* (2022), DOI: 10.1038/s41598-022-16361-6.

Paper III

Langberg, GSRE et al. "A weighted margin loss for treating imbalanced, overlapping and noisy data in cervical cancer risk prediction". *Submitted to International Journal of Medical Informatics.* (2023).

Contents

Acronyms

NCCSP Norwegian Cervical Cancer Screening Program

CRN Cancer Registry of Norway

MF Matrix factorization

GDL Geometric deep learning

HMM Hidden Markov model

HPV Human papillomavirus

CIN Cervical intraepithelial neoplasia

SIL Squamous intraepithelial lesion

LSIL Low–grade squamous intraepithelial lesions

HSIL High–grade squamous intraepithelial lesions

ASC-US Atypical squamous cells of undetermined significance

AGUS Atypical glandular cells of undetermined significance

ASC-H Atypical squamous cells cannot exclude an HSIL

ACIS Adenocarcinoma in situ

LR Logistic regression

DT Decision tree

RF Random forest

GBT Gradient boosting trees

Glossary

matrix factorization A mathematical technique for factorizing a matrix into a product of matrices.

cervix uteri/cervix The lower part of the uterus in the human female reproductive system that is in connection with the vagina.

transformation zone An area of the cervix where cervical cells are most likely to become cancerous.

human papillomavirus A group of viruses where some types are linked with cancer.

cervical carcinogenesis The transformation of normal cells into cervical cancer.

cytology Microscopic examination of a sample with exfoliated cells.

histology Microscopic examination of a tissue sample.

Bethesda system A system for reporting results from cervical cytology.

cervical intraepithelial neoplasia A system for reporting results from cervical histology.

unsupervised learning Machine learning algorithms focused on extracting intrinsic information from datasets without specific prediction targets.

supervised learning A branch of machine learning where prediction algorithms are trained on datasets with specific prediction targets.

overfitting A situation where the machine learning algorithm is unable to generalize to other data.

underfitting A situation where the machine learning algorithm is not sufficiently complex to express the information in the data.

matrix completion The task of imputing the entries of a partially observed matrix.

Chapter 1

Introduction

1.1 Motivation – more personalized cervical cancer screening recommendations

Despite being one of the most common types of female cancer worldwide [Sun+21], cervical cancer is a preventable disease [Gaf+18]. Cervical cancer gradually develops from precursor lesions in a process that may take several years and is preventable if the lesions are detected at a sufficiently early stage and managed with existing methods for clinical intervention [Sch+07b]. For the so-called early detection of cervical lesions, national screening programs have been established in developed countries. These programs recommend that each woman from the target population undergo repeated screening at regular intervals, thereby contributing to reducing cervical cancer prevalence and mortality [Vac+14].

Existing programs for cervical cancer prevention recommend repeated screening with time intervals between consecutive exams being homogeneous across the target population. For instance, the guidelines for the Norwegian Cervical Cancer Screening Program (NCCSP) currently recommend routine screening every third year for women from 25 to 33 years and every fifth year for women from 34 to 69 years [Bir22]. However, the risk of developing cervical cancer varies both with age and amongst individuals [SW13a], and the current guidelines and recommendations do not fully capture the heterogeneity of the individual risk. Studying the exam results of the NCCSP participants shows that over 65% of the women have never had abnormal results. The risk of this sub-population developing cervical cancer is thus regarded as lower than for women with several abnormal results.

Considering that the risk of developing cervical cancer varies amongst individuals, practicing fixed time intervals for regular screening causes excessive exams as this cancer only develops in a smaller segment of the target population. Differentiating recommendations based on individual risk estimates may thus improve the current mass screening strategy. The more personalized recommendations would offer, for instance, less frequent screening of middle-aged women with a series of only normal results, as they may be at considerably lower risk than adolescent women with a history of several abnormal results [SW13a]. However, one of the challenges in enabling more personalized prevention is identifying the women that may be recommended less frequent screening without compromising their protection [Sch+16a].

The maturity of the screening programs for cervical cancer prevention in Scandinavian countries creates an opportunity to study approaches to more personalized cervical cancer screening. Since the introduction of organized screening in Norway, the Cancer Registry of Norway (CRN) routinely collects

data from each examination to administer the screening program according to the national guidelines. The collected information includes the exam date, test types and corresponding test results. This data is centralized in a database containing the complete exam histories of the NCCSP population. In addition, previous studies have produced data that contains more detailed medical information about each individual but, unlike the exam history data, is not routinely collected by the CRN [Bir22]. This more detailed data include lifestyle and risk factors linked to cervical cancer development but covers only a fraction of the screening population. One approach to developing more personalized guidelines and recommendations is thus to utilize the existing databases to derive a data-driven framework for personalized cervical cancer risk assessment.

1.2 The DeCipher project for more personalized cancer screening

The DeCipher project began in 2019, funded by the Research Council of Norway and coordinated by SimulaMet in collaboration with the CRN, Lawrence Livermore National Laboratory, and Karolinska Institutet. The project promotes the idea of using data-driven methods to develop more personalized strategies in population-level cervical cancer screening. One of the primary goals of DeCipher is to explore the potential for deriving algorithms for personalized cervical cancer risk prediction.

Novelty The DeCipher follows two complementary approaches to algorithm development based on different data material. One approach combines lifestyle and risk factor information with the exam history data from the NCCSP to uncover phenotypes linked with cervical cancer development. Novel algorithms are designed to couple the static lifestyle information with the longitudinal exam histories for mining cervical cancer risk groups. The alternative direction is using only the exam history data to derive prediction algorithms for the individual risk. An assumption here is that the exam history data is sufficiently informative about the individual risk for the algorithms to distinguish between clinically different outcomes. While existing algorithms for predicting cervical cancer use more extensive data material [He+21], focusing on only the routinely collected exam history data presents a novel approach to algorithm development.

1.3 Thesis scope and research objectives

The present thesis is focusing on the direction in the DeCipher project aiming to devise algorithms that can predict the individual risk of cervical cancer development. Specifically, the scope is to prototype and evaluate data-driven prediction algorithms for cervical cancer development using only exam history data available from the CRN.

The NCCSP database includes more than 1.8 million exam histories, placing a premium on computationally efficient[1] methods. Typically, the number of exams per history is less than seven (i.e., scarce data), and the time intervals between subsequent exams are irregular. Thus, methods designed for continuous time-series data must be adapted or abandoned.

A common approach to handle scarce and irregular data is to use methods from *matrix factorization* (MF). These techniques have been used to extract latent features from longitudinal electronic medical records for downstream prediction tasks [Zho+14]. To derive prediction algorithms designed specifically for the Norwegian exam history data, this thesis leverage approaches from MF and develop methods to incorporate domain-specific aspects of the data into the latent features.

Using MF methods for latent feature extraction has also been explored within the framework of geometric deep learning (GDL) [Bro+17]. Here, machine learning algorithms are applied to data organized on similarity graphs connecting entities with similar traits, revealing structural relationships in the data [MBB17]. Via these graphs, the GDL approach incorporates structural information from the data into the latent features. Developing a GDL algorithm for the Norwegian exam history data thus necessitates similarity graphs representing the relationships between exam histories.

Considering an alternative to the MF-based methods, a recent study presented a hidden Markov model (HMM) derived from the Norwegian exam histories to simulate the transition rates between different risk categories encoded from the original exam results [Sop+20]. The HMM captures the time-varying trends observed in the exam histories in a set of parameters. Based on these parameters, an algorithm can be derived to predict cervical cancer development from individual exam histories.

However, as more than 65% of the CRN exam histories contain only normal results, the inherent data imbalance requires methods for adjusting the prediction algorithms to recognize the rare outcomes. Moreover, additional prediction difficulties may be embedded in the data, requiring specific methods to be designed for prediction accuracy.

Objectives As the scope of this thesis is to conduct an exploratory study, no formal hypotheses are assumed. To guide algorithm development and evaluation, the following research objectives are set.

1. Develop novel prediction algorithms designed for the Norwegian exam history data.

2. Evaluate the different prediction algorithms (e.g., MF, GDL and HMM) by comparing their ability to reflect the time-varying risk trends derived from data.

[1] Efficiency is understood in terms of time and memory requirements and the required number of training samples.

3. Study the accuracy of single risk estimates and unravel potential prediction difficulties embedded in the data.

This thesis encompasses three manuscripts for scientific papers addressing research objectives 1 to 3. The first manuscript focuses on objective 1, using methods from MF to develop prediction algorithms. Manuscript 2 presents prediction algorithms based on MF, HMM and GDL and uses numerical experiments to address objective 2. Finally, manuscript 3 considers objective 3.

1.4 Thesis outline

The structure of the thesis is as follows. Chapter 2 introduces background information and central concepts for the papers included in this thesis. The main topics in Chapter 2 revolve around cervical cancer development and existing prevention methods and basic concepts in machine learning. The chapter elaborates on challenges related to learning from the exam history data and how to address some of these challenges with MF techniques for latent feature extraction. Key assumptions and limitations in existing methods to consider in designing novel MF methods for the NCCSP histories are also explained. Chapter 3 delineates the narrative connecting the individual papers and their contributions, while Chapter 4 discusses the significance and impact of these papers. The terms indicated with *italic* are explained in the glossary.

Chapter 2

Background

2.1 Development of cervical cancer

The cervix uteri The lower third of the uterus in connection with the vagina is called the *cervix uteri*, from hereon referred to only as the *cervix*. The surface of the cervix is covered mainly by two types of epithelial cells; glandular cells situated in the part of the cervix leading to the uterus and squamous cells covering the mucosa in the vagina and the exterior of the cervix. The region where the glandular and squamous cells meet in the cervix is called the *transformation zone*. It is from the cells in this zone that the majority of cervical cancers develop [Bha+18].

Cervical carcinogenesis and risk factors Virtually all[1] incidences of cervical cancer are initiated after infection with *human papillomavirus* (HPV) [Wal+99]. While there are several types of HPV, estimates indicate that HPV16 and HPV18 are involved with 70% of cervical cancers [Sch+07b]. The current consensus is that persistent infection with such high-risk HPV types constitutes a necessary condition for *cervical carcinogenesis*.

Cervical carcinogenesis concerns the formation of cervical cancer from normal cells, a process which may be described by four steps: 1) HPV infection; 2) persistence of the HPV infection (development of low-grade lesion); 3) progression from a low-grade to a high-grade lesion; 4) invasive cancer. These four steps present a model for the course from persistent infection with high-risk HPV via intermediate precancerous stages to cervical cancer.

As HPV can be sexually transmitted, most individuals acquire an HPV infection over their lifetime, and the prevalence culminates in adolescence and early adulthood [SW13b]. However, most HPV infections are transient and resolve without clinical intervention. For instance, estimates suggest that 67% of HPV infections clear within one year and over 90% clear within two years [Rod+08; Sch+07a]. A malignant transformation is thus rare compared with the infection rate, indicating that HPV alone is not the cause of cervical carcinoma. On the other hand, sexual habits, long-term smoking, multi-parity and long-term use of hormonal contraceptives are potential risk factors promoting the acquisition, persistence and progression of an HPV infection [Zha+20].

Classifications Two widely established techniques for detecting cervical cancer and precursor lesions are the cervical *cytology* based on exfoliated cells and the *histology* based on a tissue sample (biopsy). The cell samples from these tests are analyzed manually in a microscope. Cytology examines individual cells

[1] More than 99% of all cervical cancers are claimed attributable to HPV infection [Wal+99]

or clusters of cells, while histology considers up to multiple cell types in the biopsy. Several terminologies and classification systems exist to report the results from these tests. Two commonly used reporting systems are the *Bethesda* and the *cervical intraepithelial neoplasia* (CIN) [Hea+06]. These systems provide consistent terminology to convey and compare results.

The Bethesda system classifies cytology results into levels of squamous intraepithelial lesion (SIL). Classifications of abnormal results include low-grade SIL (LSIL) and high-grade SIL (HSIL), indicating different levels of cell changes. Other classifications are atypical squamous cells of undetermined significance (ASCUS), atypical glandular cells of undetermined significance (AGUS), atypical squamous cells where HSIL cannot be excluded (ASC-H) and adenocarcinoma in situ (ACIS). An ASC-US result corresponds to low-grade changes, while AGUS, ASC-H and ACIS are used to denote high-grade changes [Hea+06].

For histology exams, the CIN system distinguishes between degrees of cellular changes on a scale from 1 to 3. At CIN1, the cell changes are mild (i.e., low grade) and typically regress without clinical intervention. The CIN2 changes are moderate and could require clinical action, while the CIN3 changes are severely abnormal and usually require immediate action to prevent cancer development [Hea+06]. However, regressing from the CIN2 and CIN3 stages may also occur [Ost93]. In the Bethesda system, the CIN2 and CIN3 are combined into a single group (HSIL) due to the difficulties of distinguishing these degrees of cellular changes with cytology.

Treatment and prognosis Cervical cancer can be treated through surgery, radiation, chemotherapy, or a combination of these therapies. The treatment choice depends on the cancer stage (I–IV) used to grade the extent and severity of abnormal cell changes. Detecting earlier stages of cervical cancer, followed by effective clinical management, makes it one of the most treatable forms of cancer. While treatment at an early stage can prevent cancer development, cancers detected in later stages can also be controlled with appropriate treatment and palliative care [Bha+18]. The trends from 2016 to 2020 in five-year survival of cervical cancer in Norway for women diagnosed in early stages (stage I and II) range from 80–95%. However, for more developed cancers (stage III), the five-year survival rate decreases to 54% and 20% for late-stage cancers (stage IV) [Nor21]. The high treatment success rate in the earlier stages demonstrates a potential to prevent cervical cancer with methods for early detection of disease precursors.

2.2 Cervical cancer prevention

Improved understanding of the role of HPV in the initiation and development of cervical cancer has led to various prevention methods [Gaf+18]. The two primary methods for cervical cancer prevention are HPV vaccination and routine screening. Arriving at the consensus that chronic infection with HPV causes virtually all cervical cancers, effective primary prevention relies on HPV vaccination [Chr+18].

However, the presently administered vaccines are type-specific and protect only against a few types of HPV [HD17]. Thus, vaccination against HPV provides only partial immunity and does not obviate the need for secondary prevention methods like routine screening [Fra+06].

Screening for cervical cancer prevention Cervical cancer screening aims to detect disease precursors in time for adequate treatment to prevent cancer development. The screening process generally involves a repeated examination of the target population to help determine the likely presence of cancer in possibly asymptomatic individuals. Following an HPV infection, it may take several years to develop cervical cancer. The period between infection and until the infected cells have become cancerous creates an opportunity to detect lesions early enough to prevent cancer development. Effective cervical cancer screening requires thus regular and repeated examinations to avoid late-stage treatment where the outcome is generally poorer [Bed+20].

Cervical cancer screening in Scandinavia By 1996, national programs for organized screening were implemented in the Scandinavian countries Denmark, Norway and Sweden. In the organized programs, healthcare authorities invite women from the target population to undergo routine screening according to the national guidelines [Ped+18]. One of the indicators for the success of organized cancer screening is to what extent the target population adheres to the national guidelines and is active in screening, defined as the screening coverage. In 2018, the population coverage for cervical cancer screening in the Scandinavian countries was upwards of 80% [Par+19].

The effect of national cervical cancer screening The established consensus across multiple studies is that organised cervical cancer screening in developed countries has reduced cervical cancer mortality [Jan+20]. Audits of the Swedish and Norwegian cervical cancer screening programs showed that the participating women had a reduced risk of developing cancer [And+08; NST02] and improved prognosis [And+12]. Lifestyle changes over the years have increased the exposure to HPV infections, and the prevalence and mortality of cervical cancer in Norway would most likely have been higher in the absence of screening. For instance, it has been estimated that the Norwegian program has prevented up to 68% of cervical cancers [Lön+15].

2.3 The Norwegian cervical cancer screening program

In 1995, the NCCSP was established as a central unit for nationwide screening to prevent cervical cancer in Norway. Annual reports made publicly available documents the status and development of the program. The NCCSP is administered by the CRN, aiming to reduce cervical cancer incidence and mortality in Norway. The central unit encompasses the participating women, clinicians conducting the exams, laboratories analyzing the results and the

administrative unit at the CRN. To adhere to the national guidelines, the NCCSP uses a centralized invitation procedure and issues different types of reminders to women, physicians and laboratories based on individual exam results [Bir21].

Guidelines and recommendations Since establishing the NCCSP and until 2014, all women from 25 to 69 years of age were recommended triennial screening with cytology. Testing for HPV infections was introduced in 2005 to improve follow-up on low-grade abnormalities and was incorporated into the revised guidelines in 2014. In 2018, the guidelines were further modified to account for type-specific HPV infections. A gradual transition from cytology to HPV in primary screening of 34–69 year old women began in 2019 and in 2022 this was implemented in nearly all Norwegian counties [Bir21].

As of 2022, the national guidelines in Norway recommend the women aged 25 to 33 years with a negative cytology to follow the triennial routine screening, while 34–69 year old women with a negative HPV should attend routine screening every five years. Triennial cytology is maintained for the age group 25–33 due to the high prevalence of transient HPV infections. Women detected on their primary screen with high-grade cervical lesions are advised immediate follow-up (colposcopy). Women with low-grade cervical lesions and a positive HPV test are invited to repeat the cytology and HPV test in 6–12 months, while a negative HPV test qualifies for returning to routine screening. If the repeated cytology remains low-grade and the HPV is persistently positive, the woman is recommended colposcopy, otherwise she may return to routine screening [Bir21].

Data registries Data from all cervical exams are reported to the CRN by legal obligation in the Cancer Registry Regulations. However, women with only normal results can request to be excluded from these records, which applies to 3% of the NCCSP population. To disclose the data reported to the CRN, databases for cytology, histology and HPV results were established respectively in 1991, 2002 and 2005. The CIN registry was established in 1997 and contains the data on assessment and possibly treatment of precursor lesion, while the incidence registry maintains the information about detected cancers. The NCCSP database encompasses the cytology, histology, CIN and HPV registries. By 2020, this database contained the exam histories of more than 1.8 million women attending cervical cancer screening in Norway [Bir21].

More personalized prevention Only a minority of women contracting an HPV infection will develop cervical cancer, and the risk of cancer development varies considerably among the individuals exposed to this ubiquitous infection. Utilizing information from the exam histories to estimate the prognosis for each woman can improve the risk stratification of the heterogeneous screening population. Specifically, machine learning technology developed for individual risk estimation provides information to support decisions on subsequent clinical actions. The existence and availability of NCCSP database create opportunities

for developing such technologies for more personalized clinical decision support in cervical cancer screening and improving early detection.

2.4 Concepts in machine learning

The field of machine learning has been rapidly growing in recent years, and several resources provide a comprehensive overview of how the field has evolved [Cho21; Ras15]. Machine learning commonly refers to algorithms leveraging data to perform computational tasks. Some tasks where these algorithms are used include treatment outcome prediction [Par+15], fraud detection [AAO17], speech recognition [Amo+16], and autonomous driving [Son+21]. The benefits of adopting a machine learning strategy are that larger volumes of information can be processed faster and more accurately than with manual methods. However, the accuracy of these methods relies heavily on the data used to derive the system. Particularly paramount to accurate machine learning is sufficient amounts of representative examples for how a system responds to given inputs [Sam+21].

Learning from data Machine learning may be divided into different branches distinguished by the learning strategies and applications. One of these branches is *unsupervised learning*, where algorithms extract latent information based on various assumptions about the data. These assumptions could be about the dependency structure between measurements arriving in sequence or the distribution of the measurements, whether it is balanced or skewed towards one specific outcome.

Another branch is called *supervised learning* and entails algorithms designed to learn a mapping from specific inputs to predict target outputs. For example, the goal may be to predict movie ratings from input information about personal preferences. Here, the input to the algorithm is descriptive information about the movies, often called features, organized in the form of a vector or matrix for each individual. These features are typically engineered from the data or extracted with an unsupervised method.

When learning from data in the unsupervised and supervised settings, the algorithm derives discriminative rules from the inputs through some optimization procedure, described as training or fitting. During training, the algorithm learns the appropriate rules through feedback from an objective function designed to adjust the algorithm to improve on the task. A key element in developing a machine learning algorithm is designing objective functions that capture the inherent nature of specific tasks.

Generalization When training machine learning algorithms, a cardinal objective is to derive discriminative rules that generalize across different datasets. However, a fundamental assumption in most applications is that all data presented to the algorithm originate from the same distribution. A common way to assess generalization performance in the supervised setting is to divide the available data into one set used for training and another subset for (internal)

validation. The training set is used to learn the relationships between possible inputs and corresponding examples of target outputs, and the validation set is used to assess the generalization performance during training and determine when the training should stop. Moreover, a third separate dataset may be available to test the final algorithm after training is complete. While evaluating an unsupervised algorithm can be more challenging, one approach is to combine it with an downstream supervised algorithm to assess the quality of the unsupervised solution.

Overfitting and underfitting An undesirable development that may occur during training is that the algorithm continues to improve on the training set while the performance on the validation set is decreasing. This development is referred to as *overfitting* and may occur in the supervised and unsupervised settings. Here, the algorithm may be fitting spurious relationships only found in the training data and do not generalize to other data. Possible explanations for overfitting is that there are too few examples of certain outcomes in the dataset, producing in a weaker signal, or that the algorithm is so complex that it has the capacity to model measurement noise as intrinsic information. Common methods to overcome overfitting is to impose constraints on the objective function to reduce the flexibility of the algorithm and emphasize specific aspects of the data [Ras15].

Moreover, *underfitting* is when the algorithm performs poorly on the training data as well as the validation data. Underfitting may occur if there is a low amount of data available for training and validation or if the algorithm is too simplistic to capture the main components of information [Ras15]. A potential cause for underfitting is that the objective function lacks elements to capture essential concepts within the data, such as time-dependencies. Depending on the task and available data, some algorithms may thus be more eligible than others. The aphorism "all models are wrong, but some are useful" acknowledges that prediction algorithms fall short of the complexities of reality but can nevertheless be applicable, and several different machine learning algorithms have been developed over the years; each with inherent biases [Mah20].

No free lunch An insight derived from the "no free lunch theorems" is that no single prediction algorithm is unequivocally optimal in every application [WM97]. Essential in machine learning is thus to compare different algorithms to determine which one is the more appropriate for the task. This process is often referred to as algorithm selection [Ras15]. The following briefly describes different prediction algorithms commonly used in biomedical applications.

Logistic regression Logistic regression (LR) is a supervised learning algorithm often used in classification tasks [Cox58]. To classify an input feature vector, the algorithm creates a linear combination of the features with the learned model parameters. The linear combination is then passed through an activation

function, creating the class probabilities. Usually, the input is assigned to the class corresponding to the highest probability.

Decision tree In machine learning, a decision tree (DT) is a prediction algorithm built from discriminative rules organized in a hierarchical structure[Bre+17]. Like LR, this algorithm belongs to the branch of supervised learning. The DT algorithm classifies an input by recursively querying the features according to the learned rules. The discriminative rules are inferred from the training data by optimizing the partitioning according to some metric.

Random forest The random forest (RF) algorithm consists of several DTs built separately from random subsets of the training data and organized in a parallel structure [Bre01]. The individual DTs are usually restricted to having relatively few rules. Given input data, the algorithm aggregates the output from the individual trees to determine a classification. Compared to a single decision tree, the RF can improve prediction accuracy and reduce variance in the estimates.

Gradient boosting trees Like the RF, the gradient boosting trees (GBT) algorithm builds on a collection of DTs [Fri01]. However, rather than organizing the trees in a parallel structure, GBT arranges the individual trees in a sequence. Here, the objective for each tree within the sequence is to correct the prediction errors of the predecessor tree. By combining the estimates from the individual trees as a weighted sum, the algorithm can produce a more accurate estimate than the individual trees.

Neural networks Rather than a single algorithm, neural networks describe a class of algorithms defined by a series of composite transformations of the input data [Bis94]. The transformations aim to extract discriminating features directly from the input data useful for, e.g., classification tasks. The different transformation steps refer to layers in the network, and the ordered collection of layers defines the network architecture. Different layers and architectures exist for different input formats and tasks. Like LR and DT, the multilayer Perceptron [Hay94] is suited for tabular data, while recurrent [RHW85] networks model sequential measurements, such as time series. Variants of the recurrent networks have become popular in modelling regularly sampled sequences, but these can be slow to train on large datasets and struggle to handle irregularly sampled data.

Geometric deep learning Geometric deep learning (GDL) entails neural networks applicable to data organized on structures such as similarity graphs. One of the main differences between GDL and the more traditional neural networks is thus the format of the input data. The similarity graphs organize data in a domain that is not defined by a regularly spaced grid, such as vectors and matrices, and provide more flexibility in defining the geometric properties of the

data. Examples of similarity graphs are the so-called social networks modelling social media activity or sensor networks used in communication systems. An essential step in using GDL techniques is thus to quantify the relationships in the data to define the graph structure [Bro+17].

Hidden Markov models A hidden Markov model (HMM) models a series of observations where measurements are usually restricted to depend only on the previous observation (i.e. a first-order Markov process) [BP66]. The algorithm uses a set of latent states to represent the unique elements observed in the data and a set of state transition probabilities to capture the dynamics in the observed sequences. There are several versions of the HMM, where some extend to model processes where time is not represented explicitly in discrete steps [Liu+15]. These algorithms are thus eligible for modelling longitudinal data where the measurements are irregular and sometimes scarce. While HMMs are typically unsupervised algorithms used in stochastic simulation, they can be adapted to prediction tasks.

2.5 Challenges with learning from exam history data

Canonical prediction algorithms like the LR and tree-based methods described in Section 2.4 expect sufficiently sampled and organized data. However, raw data rarely comes in the form and shape required for analysis [Tay+21]. In several situations, measurements are few and infrequent, which results in scarce and irregular datasets. Scarce data increases the chance of underfitting due to few examples to guide the learning process. Moreover, organizing irregular measurements in a uniform array for input to algorithms can dispose of important information about intrinsic structures in the data and promote underfitting [Sar21]. The combination of scarce and irregular data yields few and arbitrarily separated measurements, where essential structural information may be missing between observations.

In practice, it is also not unusual that one type of measurement occurs far more often than the others. In this case, the dataset is realized with a skewed distribution. Considering the implicit assumption of a balanced outcome representation in most prediction algorithms, highly imbalanced data can make the learning process overfit the majority outcome [Kra16].

Scarce, irregular and imbalanced exam results Throughout the NCCSP, the recommendation to commence screening at the age of 25 and cease screening at the age of 69 has remained unchanged. During these 44 years, adhering to triennial screening amounts to relatively few measurements with only 15 exams per woman [Bir22]. In practice, however, the median number of exam results per history in the population is seven.

In addition to the scarcity of results, the exams occur at erratic time intervals. While the time between consecutive exams is partially driven by the recommendations, screening occurs at the discretion of each individual.

It has been found that less than half of the NCCSP participants attended a screening at the recommended repeated intervals [Ped+17]. In combination with individual variations in the onset of an HPV infection prompting more frequent exams to follow up on abnormal results, the exam histories become irregular.

Although more than 65% of the NCCSP participants have never had an abnormal exam result, the current recommendation is a regular examination of the whole population to protect the sub-population of high-risk individuals. Consequently, the number of normal results exceeds the number of abnormal results. Specifically, more than 90% of the total results are normal, and less than 3% of the results are high-grade, materializing in a skewed distribution of exam results. This level of imbalance in the results can be detrimental to prediction performance [Jap00].

Accuracy in exam results The exam history data contained in the NCCSP database include only the results from conventional tests (i.e., cytology, histology and HPV) [Bir22]. Each test type associates some level of uncertainty to correctly indicating an abnormal (sensitivity) or a normal (specificity) result. Poor test sensitivity will lead to false negatives, while poor specificity yields false positives.

Estimates on the sensitivity and specificity of the tests used in cervical cancer screening vary considerably between studies. A review study summarized that the sensitivity and specificity reported for cervical cytology varied from 30 to 87% and 86 to 100%, respectively [Nan+00]. Compared to cytology, the HPV test typically has higher sensitivity [May+07], although the specificity can suffer from false positives caused by the frequently occurring transient infections [Cuz+06].

The phrase "garbage in, garbage out" is a classic saying in computer science relating errors in the input data to errors in the algorithm outputs. Besides limited accuracy in cervical exam tests, inadequate sampling, sample preparation, and the interpretation of findings may also bias the results. The effect of these distortions is noisy and ambiguous data that can degrade algorithm accuracy.

2.6 Feature learning from scarce, irregular and imbalanced data with matrix factorization

A principal consideration when selecting an algorithm for a practical application is whether it accommodates the input format of the data. A popular class of algorithms for scarce and irregular data is MF methods. These algorithms create lower dimensional embeddings of the data for dimension reduction and latent feature representations. The features can be utilized in downstream prediction tasks [Zho+14] and to derive recommender systems [KBV09]. Techniques from recommender systems became especially popular for movie rating prediction during the Netflix Prize competition. The challenge in this competition was to improve the current recommendation system using data on the movie ratings provided by different users [BL+07]. In this data, each user had typically rated only a few different movies, resulting in scarce and irregular measurements.

Representing scarce and irregular data One popular way to represent the Netflix data is as a matrix $\boldsymbol{X} \in \mathbb{R}^{N \times M}$. Here, the $1 \leq n \leq N$ rows indicate different users and the $1 \leq m \leq M$ movies are on the columns. A movie rating is a numerical value $X_{n,m} > 0$ in the matrix entries (n, m), while zero entries indicate unrated movies. As each user n provides only a few ratings for different movies, most entries of the individual rating profiles $\boldsymbol{X}_n$ are zero, with arbitrary numbers of unrated movies between the ratings. By representing $\boldsymbol{X}$ as a low-rank factor model, a complete set of latent features can be derived from the scarce data matrix and utilized in downstream prediction tasks. However, an underlying premise for the low-rank factor model is that the lower dimensional representation suffices to capture the main aspects of the data [BK07].

A low-rank factor model In the Netflix data, different users are found to share movie preferences and exhibit similar rating behaviour. This observation implies that the rating profiles $\boldsymbol{X}_n$ are correlated and that the variability in $\boldsymbol{X}$ may be well described by only a few, i.e., $r \ll \min\{N, M\}$ elementary components rather than in the original dimensions of $\boldsymbol{X}$. An alternative representation of $\boldsymbol{X}$ is thus as a low-rank factor model where $\boldsymbol{X}$ is decomposed into two factors $\boldsymbol{U} \in \mathbb{R}^{N \times r}$ and $\boldsymbol{V} \in \mathbb{R}^{M \times r}$. Here, each movie rating is modelled as a linear combination $X_{n,m} \approx \boldsymbol{U}_n \boldsymbol{V}_m^\top$, where $\boldsymbol{U}_n$ are the user-specific coefficients and $\boldsymbol{V}_m$ is a linear predictor for the columns of $\boldsymbol{X}$ (i.e., the different movies).

The latent features Assuming r elementary components in $\boldsymbol{X}$ translates to r latent features in the factor matrices. In the context of the Netflix data, each row of $\boldsymbol{U}$ indicates the preferences of a user with a weight vector for the latent features. The features can have interpretations such as movie genre, the actors involved and the time of the release, and the weights reflect their relative importance to each user. The interpretation of $\boldsymbol{V}$ is the information about the latent features shared between the users.

In big-data paradigms, a low-rank factor model is more memory-efficient since $\boldsymbol{U}$ and $\boldsymbol{V}$ consume less computer disk space than $\boldsymbol{X}$ due to their lower dimensionality. Determining the size of the factor dimension is usually approached with heuristic methods to optimize the model to the data. Moreover, the choice of r impacts the complexity of the data model. Increasing r leads to more latent features and is more prone to overfitting while choosing r too small can result in underfitting.

Estimating factor matrices from imbalanced data By employing techniques from *matrix completion*, the factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ can be estimated from the rated entries in $\boldsymbol{X}$. A typical approach is to minimize the overall discrepancy $\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^\top$ between the known $\boldsymbol{X}$ and the estimated $\boldsymbol{U}\boldsymbol{V}^\top$ entries, as quantified by the Frobenius norm $\|\cdot\|_F$ [NKS19]. When $\boldsymbol{X}$ is scarce, the relevant entries to minimize over in the residual matrix can be specified by element-wise

multiplication with an indicator matrix $\boldsymbol{W}$. This yields the optimization objective

$$\min_{\boldsymbol{U},\boldsymbol{V}} \frac{1}{2} \left\| \boldsymbol{W} \odot (\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^\top) \right\|_F^2 . \tag{2.1}$$

In the indicator matrix matrix, $W_{n,m} = 1$ if $X_{n,m}$ holds a non-zero value and otherwise is zero. To account for the data imbalance due to some values occurring far more often than others, the indicator matrix may be replaced by a weighting matrix [SJ03].

In the weight matrix, $W_{n,\tilde{m}} \gg W_{n,m}$ if the entity m is more common than entity $\tilde{m}$. This technique has the effect of emphasizing the signal from rare entries that may otherwise be suppressed by the more frequently occurring entries. On the contrary, entries that should not be emphasized due to high uncertainty tied to the underlying measurement can be down-weighted to reduce the influence of this observation in training. In general, the design of weight matrix $\boldsymbol{W}$ depends on the data and the specific application [Sch+16b].

Using methods of alternating minimization, problems like (2.1) can be solved efficiently in both time and memory requirements [JNS13]. These algorithms achieve an approximate solution to the original problem by alternating between minimizing with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$. Central to matrix completion is deriving factor estimates that captures the intrinsic structures in $\boldsymbol{X}$. To potentially improve the generalization of the solution and reduce overfitting, constraints can be added to (2.1). Constraints may also be used to encourage specific properties of the factor matrices to prevent underfitting.

Improving factor estimates with regularisation Constraining the factor matrices by incorporating domain-specific properties can improve their generalization abilities as downstream features and their fit to the data. One way to implement a wide range of constraints is by adding regularization terms to the objective formulation [KM10]. Although the choice of regularisers is problem-specific, some techniques are widely used across applications. For instance, L2 regularisation on the factor matrices can prevent overfitting by controlling the complexity of the factors [MS07]. Including L2 regularisation, the objective in (2.1) extends to

$$\min_{\boldsymbol{U},\boldsymbol{V}} \left\{ \frac{1}{2} \left\| \boldsymbol{W} \odot (\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^\top) \right\|_F^2 + \alpha_1 \left\| \boldsymbol{U} \right\|_F^2 + \alpha_2 \left\| \boldsymbol{V} \right\|_F^2) \right\} \tag{2.2}$$

Larger $\alpha > 0$ in (2.2) will reduce the chance of overfitting to the non-zero entries in $\boldsymbol{X}$ by penalizing large elements of the factor matrices.

Interlude: Time-varying measurements In some applications, the columns of the data matrix $\boldsymbol{X}$ may not be representing independent entities such as movies but rather a structural relationship like time dependencies. Considering scarce measurements that are collected irregularly over time, these may still be fitted into a matrix $\boldsymbol{Y} \in \mathbb{R}^{N \times T}$. Here, each row $\boldsymbol{Y}_n \in \mathbb{R}^T$ represents a longitudinal vector with measurements organized over $1 \leq t \leq T$ time points.

This data representation preserves information about the temporal ordering of the measurements. Assuming $\boldsymbol{Y}$ is of low-rank factor, each profile vector can be decomposed into a set of vector-specific coefficients $\boldsymbol{U} \in \mathbb{R}^{N \times r}$ and shared time-varying basic profiles $\boldsymbol{V} \in \mathbb{R}^{T \times r}$.

Modelling longitudinal data Applications of MF to time-dependent data remain scarce, possibly due to the added difficulty of inferring and fitting the temporal dynamics from only a few available measurements. In densely sampled time series data, dependencies have been encoded using similarity graphs, linking together data points in time [YRD16]. However, in longitudinal data where measurements are more scarce, the accuracy of the similarity graphs may suffer from difficulties in quantifying profile similarities.

As an alternative to graph-based approaches, the temporal structure from scarcely sampled data can be encoded as regularizers. One strategy is to assume temporal smoothness in the longitudinal basic profiles $\boldsymbol{V}$ and apply a finite difference matrix $\boldsymbol{D} \in \{-1, 0, 1\}^{T \times T}$ to reveal the smoothness over adjacent time points [Zho+11]. Taking the Frobenius norm over the matrix product $\boldsymbol{DV}$ quantifies the overall level of irregularity, which can be minimized with the objective

$$\min_{\boldsymbol{U},\boldsymbol{V}} \left\{ \frac{1}{2} \left\| \boldsymbol{W} \odot (\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V}^\top) \right\|_F^2 + \alpha_1 \left\| \boldsymbol{U} \right\|_F^2 + \alpha_2 \left\| \boldsymbol{V} \right\|_F^2 + \alpha_3 \left\| \boldsymbol{DV} \right\|_F \right\} \quad (2.3)$$

Here, the term $\left\| \boldsymbol{DV} \right\|_F$ will promote smoothness in $\boldsymbol{V}$, making the basic profiles vary more slowly. The regularization induced by $\boldsymbol{DV}$ is uniform across the measurements in time, penalizing all deviations at neighboring time points equally much. However, a more flexible algorithm may be required to prevent underfitting in applications where the temporal dynamics varies at a non-uniform rate.

2.7 Summary

Cervical cancer is a disease that gradually develops over time from a region in the lower third of the uterus. Developing abnormal cervical cells follows from a persistent infection with high-risk HPV, progressing via pre-cancerous lesions to invasive cancer. Existing methods for detecting cervical cancer and the pre-cancerous states are based on cytology, histology and HPV tests. These methods are widely adopted in national programs for cervical cancer screening in Scandinavian countries. These screening programs recommend that individuals from the target population undergo regular and repeated examination to detect disease precursors in time for adequate treatment and cancer prevention.

The cervical cancer screening program in Norway (NCCSP) recommends regular screening every three to five years for women from 25 to 69 years of age. The program has been demonstrated to reduce cervical cancer incidence and mortality in the NCCSP population. Data from all the cervical exams are registered in the NCCSP database at the Cancer Registry of Norway (CRN),

containing the complete exam histories of more than 1.8 million women. The CRN uses information from the exam history data to administer the NCCSP according to the national guidelines. The existence and availability of the NCCSP data create opportunities for devising machine learning technologies to develop more personalized guidelines, adapting the recommendations to the individual risk for more targeted screening.

Machine learning entails algorithms that can learn a mapping from input data to specific target outputs to predict future events. Several prediction algorithms have been developed for sufficiently sampled and organized data. However, the NCCSP exam histories are scarce and irregular in the sense that they contain relatively few measurements at irregular time intervals. Applying standard machine learning methods to this data increase the chance of underfitting due to lack of examples. Moreover, as the vast majority of exam results are normal and cervical cancer only develops in a smaller segment of the screening population, this create a data imbalance that promotes overfitting. Additional learning difficulties may arise from the less than perfect sensitivity and specificity in the exam tests, potentially biasing the results.

Popular algorithms for scarce and irregular data are based on techniques from matrix factorization. These algorithms are commonly used to learn a latent feature representation by creating lower dimensional embeddings of the original data. The basic assumption for this approach is that the information from different entities are correlated such that the main aspects of the data can be captured with a lower dimensional representation. Intrinsic structures observed in the Norwegian exam history data include local variability about abnormal results and temporal shifts in the times when these results are first detected. To properly encode this information in the latent features used for downstream prediction tasks calls for development of novel algorithms designed for the NCCSP exam history data.

Chapter 3

Summary of Papers

This thesis draws on three original manuscripts for scientific papers that have been submitted to or published in peer-reviewed journals. The papers aim to address the research objectives in Section 1.3 and are presented here in the order of writing. The first paper presents prediction algorithms based on MF methods designed to capture the intrinsic structures in the Norwegian exam history data; the second paper compares the calibration accuracy in different prediction algorithms to the the time-varying absolute risk of cervical cancer derived from data; whereas the third and final paper studies confidence and correctness in single predictions to elucidate on potential difficulties embedded in the data.

Paper I Matrix factorization for the reconstruction of cervical cancer screening histories and prediction of future screening results

Developing a prediction algorithm for the Norwegian exam history data presents several technical challenges to algorithm design. Methods based on MF can accommodate scarcity and irregularity in the data but needs further development to capture the intrinsic structures in the NCCSP exam history data. For instance, longer time intervals with normal results are typically followed by shorter intervals with repeated exams after detecting an abnormal result. Sequences like this are found in several histories and reveal correlations in the data, but the onset of these sequences varies by different time delays. The developed algorithm should therefore allow for larger local variability and temporal shifts between correlated exam histories.

This paper presents prediction algorithms based on techniques from MF designed for the NCCSP data. The developed MF techniques extract structural information from the exam histories to derive two factor matrices for downstream prediction of cervical cancer development. Specifically, a regularization term is introduced to model the local variability in the temporal structure observed around abnormal exam results. Moreover, a discrepancy term is proposed to adjust the shared basic profiles to the delay in each history, increasing the correlation between shifted histories. The discrepancy term also includes a weight matrix designed to accommodate the imbalance and uncertainty in exam results. The weights incorporate estimates for the uncertainty in the cervical exam types and the importance of the individual exam results. In the downstream task of classifying exam results, a method is presented to adjust the classification threshold to the imbalance in the data.

The paper studies the proposed algorithms in numerical experiments on synthetic data and NCCSP exam histories. By generating synthetic data resembling the exam histories, ground truths are available and used to evaluate the accuracy of the factor estimates and indicate useful regularisers and discrepancy terms for fitting the data. In experiments on NCCSP data, the ability of the algorithms to predict exam results is demonstrated by measuring the similarity between Kaplan-Meier curves derived from hold-out data and predictions.

Paper II Towards a data-driven system for personalized cervical cancer risk stratification

An interpretation of the "no free lunch theorems" in machine learning is that no single prediction algorithm is superior in every application. Comparing different algorithms will thus underline which category of methods is more suited to a specific problem. Besides MF, algorithms predicting the future risk and exam results in the Norwegian exam history data can also be derived with GDL and a HMM. However, the imbalance in the data may drive the prediction estimates towards a normal result (majority outcome) and make the algorithm fail to detect the rare high-risk outcomes. Moreover, changes in the screening guidelines and lifestyle habits over time have potentially altered this distribution, resulting in a non-stationary environment.

This paper studies the calibration accuracy in different prediction algorithms by comparing their output estimates against exam results from the data over time. Algorithms based on MF, GDL and HMM are adapted to incremental learning by efficiently updating their current beliefs using all data up to the previous prediction step, making more informed predictions over time. To avoid overfitting, the algorithm outputs are adjusted to the imbalance in the exam results. However, analyzing the distribution of results in the NCCSP data shows that it is also non-stationary, as the proportion of exam results changes over time. Adapting to the observed drift and imbalance, a time-inhomogeneous classifier is created to predict the next results over segments of age intervals.

In numerical experiments, prediction algorithms for time-varying data based on MF, GDL and HMM are studied together with LR, RF and GBT algorithms designed for static data. Evaluating the impact of data imbalance on algorithm inference shows that the skew in the exam results drive predictions towards normal. Compared to using a time-homogeneous classifier, the proposed time-inhomogeneous strategy for predicting non-stationary and imbalanced data improves the prediction accuracy. Finally, the calibration accuracy is evaluated by comparing absolute risk curves derived from algorithm predictions and hold-out data. The risk curves derived from the MF algorithm show the closest resemblance to the hold-out data. A close resemblance between these curves indicates that prediction

algorithms can accurately reflect the estimated event rates for a group of Norwegian women.

Paper III A weighted loss function for treating imbalanced, overlapping and noisy data in cervical cancer risk prediction

Absolute risk curves derived from aggregated statistics give limited insights into the accuracy of prediction algorithms. That is, promising prediction performance at the cohort level does not necessarily translate to high accuracy for individual predictions. Visual inspection of exam histories in the NCCSP data reveals a close resemblance between histories of contrasting clinical significance. Together with the imbalance and the less than perfect accuracy in exam results, this overlap in the historical information leading up to differently defined endpoints adds to the challenge of deriving accurate personalized prediction algorithms.

This paper introduces a probabilistic margin score to measure the confidence and correctness of individual probabilistic algorithm predictions. The information from this score is utilized in a weighted loss function, maximizing prediction margins to improve generalization when training machine learning algorithms in overlapping data. Moreover, incorporating adaptive weights into this loss adjusts the optimization to focus on the challenging data. However, as these weights are typically based on the perceived prediction difficulty, the accuracy of the estimates may suffer from noisy data. Smooth adjustment to the weights is therefore introduced to constrain the updates by aggregating information over the iterates.

Results from numerical experiments on synthetic data and exam history data from the NCCSP indicate that smooth weights are more robust towards the perturbation of the target label and can improve prediction accuracy on noisy data. Furthermore, results suggest that losses based on margin maximization can improve generalization in overlapping data. Studying prediction confidence and correctness via the coverage profiles introduced in this paper suggests that the majority of prediction difficulties arise from resembling histories moving into different outcomes.

Chapter 4

Discussion

This thesis explores the potential for deriving accurate machine learning methods to estimate cervical cancer development by prototyping and evaluating prediction algorithms using only the exam history data from the NCCSP database. Previous studies on developing cervical cancer prediction algorithms have used more extensive data with lifestyle and risk factor information, going beyond the data routinely collected by the CRN [He+21]. Currently, this information is unavailable for all individuals in the NCCSP population and algorithms derived from only the exam history data may thus be more easily integrated into the existing screening program than algorithms derived from more comprehensive data material. However, central to this approach is that the Norwegian exam history data is sufficiently informative for individual prediction. Moreover, to extract predictive information from the scarce and irregular longitudinal measurements of noisy and imbalanced exam results, novel algorithms must be designed for the NCCSP exam histories.

4.1 Re-visiting the research objectives

This thesis comprises three research objectives to guide the steps in algorithm development. The first objective is to derive novel prediction algorithms designed for the Norwegian exam history data; the second objective is to compare the calibration of different prediction algorithms against the time-varying risk of cervical cancer as derived from data; and the third and final objective is to study the accuracy in single risk estimates and potential prediction difficulties embedded in the data to guide further algorithm development.

Prediction algorithms based on MF methods designed for the Norwegian exam history data are presented in Paper I to address the first research objective. These methods extract latent features that, together with input information from a specific history, are used for predictions. The paper introduces a regularization term for the time-varying basic profiles capturing the shared trends in the data to incorporate temporal structure information in the latent features. In line with the approach of [Zho+14], the regularization assumes that the basic profiles are smooth in time but allows for more local variability, as observed from the history data. The specific regularization model is chosen to reduce the penalization of the profiles at faster scales, but the approach generalizes to other applications of non-uniform smoothness constraints.

Furthermore, Paper I introduces a discrepancy term to enhance the correlation between structurally similar exam histories having different time delays prior to the first abnormal result. The observed delays may be explained by variations in the onset of a persistent HPV infection, resulting in variable time intervals before

detecting the abnormalities. The discrepancy term aligns the basic profiles to each history to increase emphasis on the common structures in the data. This alignment strategy can be used to enhance correlations in data from other applications subject to local shifts.

Using techniques from MF is not the only way to derive prediction algorithms designed for the NCCSP exam histories. Comparing the prediction accuracy of alternative approaches will highlight promising directions for further development. The second research objective is considered in Paper II, comparing time-varying absolute risk estimates derived from hold-out data and different prediction algorithms, including MF, GDL and HMM. Predicting results from the Norwegian exam history data over time requires adapting these algorithms to incremental learning. On the other hand, the underlying temporal dependencies between exam results may be obscured by measurement scarcity. Thus, prediction algorithms not typically used with longitudinal data (i.e., LR, GTB and RF) are also included in the comparison. This latter group of algorithms is used to investigate if time-invariant methods underfit compared to time-dependent algorithms.

Using absolute risk curves in Paper II to represent prediction accuracy over time show that the different prediction algorithms can reflect the trends observed from data. This result demonstrates a potential for using data-driven methods to predict individual cervical cancer development from only the current exam history. The main body of existing works on developing prediction algorithms for cervical cancer development is focused on classifying the risk of cervical cancer development, using measurements from only a single point in time as input[Cur21; Rot+18]. However, results from Paper II suggest that utilizing information about the longitudinal trends in the data can be beneficial to prediction accuracy. Here the algorithms designed for time-varying data (MF, GDL and HMM) gave more accurate absolute risk estimates than the alternative group of methods expecting static input data (LR, GTB and RF). Moreover, amongst the time-varying algorithms, the MF and HMM were found to be superior to the GDL.

The third and final research objective related to the accuracy of single predictions and potential prediction difficulties in the exam history data is studied in Paper III. A basic assumption in Paper I and Paper II is that the main reason for prediction difficulties is the imbalance in exam results. However, even with a significant disproportion in prediction targets, canonical algorithms may still give accurate estimates if the data distributions are separable [BPM05]. On the other hand, a class overlap is not uncommon in imbalanced data and can lead to severe learning difficulties [BPM04]. In contrast to the strategies for alleviating only the class imbalance in Paper I and Paper II, this paper presents methodology that anticipates both information overlap and imbalance in the data.

The results from numerical experiments in Paper III indicate that the proposed strategy can correctly detect high-grade cervical lesions (minority outcome) in several different exam histories. However, supplementary analyses show that the majority of prediction errors come from high-grade histories with a strong resemblance to normal histories. Identifying this overlap expands the

understanding of the prediction difficulties embedded in the Norwegian exam history data. Addressing this challenge is considered a topic for future work.

4.2 Limitations and future work

Numerical results from Paper II and Paper III presented in this thesis demonstrate a potential for using data-driven algorithms to predict cervical cancer development from individual exam histories. The similarity in cervical cancer screening guidelines and data collection practices within Nordic countries creates opportunities to assess whether the prediction algorithms can generalize beyond the Norwegian population. However, the algorithms presented in this thesis have not been evaluated in external populations due to limited data availability.

Previous studies on cervical cancer prediction algorithms may have suffered from scarce data [MRA+21] and uneven proportions of the target variable outcome [Yan+19]. Although the MF algorithm introduced in Paper I can fit the scarce, irregular and imbalanced exam results, the latent features may be improved by addressing some of the current limitations to the technical algorithm specifications. For instance, the algorithm considers only one exam result per visit, while there may occasionally be multiple results from the same date. One direction for future work is thus to extend the MF framework to utilize information from all the results in an exam history.

The HMM algorithm in Paper II improves technically on this aspect of the MF by accounting for an arbitrary number of exam results per visit as well as considering the exam types. However, both the HMM and MF are by design focused on the more holistic trends in the data, while capturing more of the local structures may important to improve discrimination. For instance, the top performing methods in the Netflix Prize competition utilized a complimentary set of algorithms, combining MF with techniques from nearest neighbour approaches [Tak+08]. In combination, these algorithms can be used to merge information from different levels of structure in the data as MF uncovers the more global trends and neighbourhood algorithms find more localized relationships. Extracting more localized structures can help differentiate between histories sharing several similarities, as described in Paper III. However, a challenge relating to the neighbour approaches is quantifying the similarity between the exam histories due to the measurement scarcity and irregularity.

Quantitatively comparing exam histories is also a challenge encountered with the GDL algorithm in Paper II. The GDL algorithm combines principles from MF with similarity graphs derived from neighbourhood methods to facilitate constraints on the latent features. In these graphs, the connected nodes represent resembling histories as estimated by some similarity measure. Methods to accurately quantify their resemblance can reveal minor differences in overlapping histories leading to clinically different endpoints. However, applying standard measures like the Euclidean distance to the scarce and irregular exam histories may yield inaccurate results and degrade performance [Gog+21]. Thus,

developing a measure for the affinity between histories will enable using more localized methods.

Re-visiting the HMM approach, a hierarchical HMM was recently proposed to model cervical cancer development, assuming only some portion of the screening population were at risk of cancer development [Men+22]. Recall that in the NCCSP data, the majority of the women have only normal results, while a smaller portion of the population develop low-grade lesions and only a few of these progress to high-grade where some might develop cancer. The proposed hierarchical HMM distinguishes between women susceptible to processing from a persistent HPV infection and where the infection will resolve without any clinical intervention. Rather than using one set of parameters as in [Sop+20], this algorithm has increased modelling capacity by using separate parameters for each category. The hierarchical HMM thus present another direction to potentially remedy the information overlap in the exam histories by addressing different level of structure in the data.

Other approaches to handle problems of information overlap co-occurring with data imbalance may involve algorithms for synthetic data generation [VEP21]. Combining empirical data with synthetic examples can increase the visibility of rare and challenging samples, creating a more balanced and informative dataset for algorithm training. Here, the existing HMMs from [Sop+20] and [Men+22] can be used as stochastic simulators to produce synthetic data resembling the NCCSP exam histories and augment the original training data.

While approaches to address the issues of information overlap may rely on continued algorithm development and synthetic data generation, another direction is to derive an alternative representation of the encoding of results to enhance the dissimilarity between histories. Throughout the papers presented in this thesis, the original results reported based on the Bethesda and CIN systems are compressed into four categories, similar to [Sop+20]. The simplified results are aligned with the four-stage model for cervical carcinogenesis to represent the state of each woman. Compressing the original results can reduce inaccuracies and noise in individual results but may also dispose of discriminative information about individual risk. Using too few categories can therefore make the exam histories seem more alike. On the other hand, using too many categories may increase the scarcity per category and potentially increase the noise. Directions for future work thus involve exploring the number of categories and alternative data representations to enhance the predictive information in the data.

4.3 Summary and conclusions

The availability of population-level data on cervical exams routinely collected and centrally stored in a national cancer registry create opportunities to derive prediction algorithms for the individual risk of cervical cancer development. Such algorithms can be used to develop more personalized screening guidelines and recommendations for more targeted cervical cancer prevention. Here, a pivotal question is if the NCCSP exam history data is sufficiently informative

for deriving algorithms with satisfactory accuracy in individual predictions. Moreover, algorithms need to be designed specifically to accommodate the intrinsic structures of this data. This methodology may also be applicable to data from other cancer types, such as colorectal, breast and prostate cancer.

This thesis presents novel approaches to predicting cervical cancer development from individual exam histories by exposing methodology developed specifically for the NCCSP data. The ability of the algorithms to predict cervical cancer development from only the exam history data is demonstrated through numerical experiments. The results indicate a potential for using data-driven algorithms to derive guidelines differentiating according to the individual risk for more personalized recommendations. Moreover, comparing several different algorithms suggest that approaches based on MF and HMM are promising directions for personalized risk prediction.

However, to improve the accuracy of algorithm predictions, difficulties arising from an information overlap within the data must be overcome. The overlap appears as similarities between exam histories leading to endpoints of different clinical relevance, and recognizing it improves the understanding of the complexity within the NCCSP data. Approaches to remedy overlapping data may involve further algorithmic development to focus on more localized structures in the data besides the more holistic perspective. Alternatively, supplementing the training data with synthetic data or adopting a different representation of the exam results may increase the separability of the exam histories. Aligning approaches to increase the predictive information in the data with the algorithms presented in this thesis is expected to improve their prediction accuracy.

Bibliography

[Amo+16] Amodei, D. et al. “Deep speech 2: End-to-end speech recognition in english and mandarin”. In: *International conference on machine learning*. PMLR. 2016, pp. 173–182.

[And+08] Andrae, B. et al. “Screening-preventable cervical cancer risks: evidence from a nationwide audit in Sweden”. In: *Journal of the National Cancer Institute* vol. 100, no. 9 (2008), pp. 622–629.

[And+12] Andrae, B. et al. “Screening and cervical cancer cure: population based cohort study”. In: *Bmj* vol. 344 (2012).

[Bed+20] Bedell, S. L. et al. “Cervical cancer screening: past, present, and future”. In: *Sexual medicine reviews* vol. 8, no. 1 (2020), pp. 28–37.

[Bha+18] Bhatla, N. et al. “Cancer of the cervix uteri”. In: *International journal of gynecology & obstetrics* vol. 143 (2018), pp. 22–36.

[Bir21] Birgit Engesæter Linn Fenna Groeneveld, G. B. S. o. A. T. *Annual report 2022 – Screening activity and results from the Cervical Program (original title: Årsrapport 2020 – Screeningaktivitet og resultater fra Livmorhalsprogrammet)*. Cancer Registry of Norway – Institute for population-based cancer research, 2021.

[Bir22] Birgit Engesæter Linn Fenna Groeneveld, G. B. S. o. A. T. *Annual report 2020 (in Norwegian)*. Cancer Registry of Norway, 2022.

[Bis94] Bishop, C. M. “Neural networks and their applications”. In: *Review of scientific instruments* vol. 65, no. 6 (1994), pp. 1803–1832.

[BK07] Bell, R. M. and Koren, Y. “Lessons from the netflix prize challenge”. In: *Acm Sigkdd Explorations Newsletter* vol. 9, no. 2 (2007), pp. 75–79.

[BL+07] Bennett, J., Lanning, S., et al. “The netflix prize”. In: *Proceedings of KDD cup and workshop*. Vol. 2007. Citeseer. 2007, p. 35.

[BP66] Baum, L. E. and Petrie, T. “Statistical inference for probabilistic functions of finite state Markov chains”. In: *The annals of mathematical statistics* vol. 37, no. 6 (1966), pp. 1554–1563.

[BPM04] Batista, G. E., Prati, R. C., and Monard, M. C. “A study of the behavior of several methods for balancing machine learning training data”. In: *ACM SIGKDD explorations newsletter* vol. 6, no. 1 (2004), pp. 20–29.

[BPM05] Batista, G. E., Prati, R. C., and Monard, M. C. “Balancing strategies and class overlapping”. In: *International symposium on intelligent data analysis*. Springer. 2005, pp. 24–35.

[Bre+17] Breiman, L. et al. *Classification and regression trees*. Routledge, 2017.

[Bre01] Breiman, L. "Random forests". In: *Machine learning* vol. 45, no. 1 (2001), pp. 5–32.

[Bro+17] Bronstein, M. M. et al. "Geometric deep learning: going beyond euclidean data". In: *IEEE Signal Processing Magazine* vol. 34, no. 4 (2017), pp. 18–42.

[Cho21] Chollet, F. *Deep learning with Python*. Simon and Schuster, 2021.

[Chr+18] Chrysostomou, A. C. et al. "Cervical cancer screening programs in Europe: the transition towards HPV vaccination and population-based HPV testing". In: *Viruses* vol. 10, no. 12 (2018), p. 729.

[Cox58] Cox, D. R. "The regression analysis of binary sequences". In: *Journal of the Royal Statistical Society: Series B (Methodological)* vol. 20, no. 2 (1958), pp. 215–232.

[Cur21] Curia, F. "Cervical cancer risk prediction with robust ensemble and explainable black boxes method". In: *Health and Technology* vol. 11, no. 4 (2021), pp. 875–885.

[Cuz+06] Cuzick, J. et al. "Overview of the European and North American studies on HPV testing in primary cervical cancer screening". In: *International journal of cancer* vol. 119, no. 5 (2006), pp. 1095–1101.

[Fra+06] Franco, E. L. et al. "Issues in planning cervical cancer screening in the era of HPV vaccination". In: *Vaccine* vol. 24 (2006), S171–S177.

[Fri01] Friedman, J. H. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[Gaf+18] Gaffney, D. K. et al. "Too many women are dying from cervix cancer: Problems and solutions". In: *Gynecologic oncology* vol. 151, no. 3 (2018), pp. 547–554.

[Gog+21] Gogineni, V. C. et al. "Data-driven personalized cervical cancer risk prediction: A graph-perspective". In: *2021 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE. 2021, pp. 46–50.

[Hay94] Haykin, S. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[HD17] Harper, D. M. and DeMars, L. R. "HPV vaccines–a review of the first decade". In: *Gynecologic oncology* vol. 146, no. 1 (2017), pp. 196–204.

[He+21] He, B. et al. "Prediction models for prognosis of cervical cancer: systematic review and critical appraisal". In: *Frontiers in public health* (2021), p. 398.

[Hea+06] Health, W. H. O. R. et al. *Comprehensive cervical cancer control: a guide to essential practice*. World Health Organization, 2006.

[Jan+20] Jansen, E. E. et al. “Effect of organised cervical cancer screening on cervical cancer mortality in Europe: a systematic review”. In: *European Journal of Cancer* vol. 127 (2020), pp. 207–223.

[Jap00] Japkowicz, N. “The class imbalance problem: Significance and strategies”. In: *Proc. of the Int’l Conf. on Artificial Intelligence*. Vol. 56. Citeseer. 2000, pp. 111–117.

[JNS13] Jain, P., Netrapalli, P., and Sanghavi, S. “Low-rank matrix completion using alternating minimization”. In: *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. 2013, pp. 665–674.

[KBV09] Koren, Y., Bell, R., and Volinsky, C. “Matrix factorization techniques for recommender systems”. In: *Computer* vol. 42, no. 8 (2009), pp. 30–37.

[KM10] Keshavan, R. H. and Montanari, A. “Regularization for matrix completion”. In: *2010 IEEE International Symposium on Information Theory*. IEEE. 2010, pp. 1503–1507.

[Kra16] Krawczyk, B. “Learning from imbalanced data: open challenges and future directions”. In: *Progress in Artificial Intelligence* vol. 5, no. 4 (2016), pp. 221–232.

[Liu+15] Liu, Y.-Y. et al. “Efficient learning of continuous-time hidden markov models for disease progression”. In: *Advances in neural information processing systems* vol. 28 (2015).

[Lön+15] Lönnberg, S. et al. “Cervical cancer prevented by screening: Long-term incidence trends by morphology in Norway”. In: *International journal of cancer* vol. 137, no. 7 (2015), pp. 1758–1764.

[Mah20] Mahesh, B. “Machine learning algorithms-a review”. In: *International Journal of Science and Research (IJSR).[Internet]* vol. 9 (2020), pp. 381–386.

[May+07] Mayrand, M.-H. et al. “Human papillomavirus DNA versus Papanicolaou screening tests for cervical cancer”. In: *New England Journal of Medicine* vol. 357, no. 16 (2007), pp. 1579–1588.

[MBB17] Monti, F., Bronstein, M., and Bresson, X. “Geometric matrix completion with recurrent multi-graph neural networks”. In: *Advances in neural information processing systems* vol. 30 (2017).

[Men+22] Meng, R. et al. “Hierarchical continuous-time inhomogeneous hidden Markov model for cancer screening with extensive followup data”. In: *Statistical Methods in Medical Research* vol. 31, no. 12 (2022), pp. 2383–2399.

[MRA+21] Mehmood, M., Rizwan, M., Abbas, S., et al. “Machine learning assisted cervical cancer detection”. In: *Frontiers in Public Health* (2021), p. 2024.

[MS07] Mnih, A. and Salakhutdinov, R. R. "Probabilistic matrix factorization". In: *Advances in neural information processing systems* vol. 20 (2007).

[Nan+00] Nanda, K. et al. "Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: a systematic review". In: *Annals of internal medicine* vol. 132, no. 10 (2000), pp. 810–819.

[NKS19] Nguyen, L. T., Kim, J., and Shim, B. "Low-rank matrix completion: A contemporary survey". In: *IEEE Access* vol. 7 (2019), pp. 94215–94237.

[Nor21] Norway, C. R. of. *Cancer in Norway 2020 - Cancer incidence, mortality, survival and prevalence in Norway*. Cancer Registry of Norway – Institute for population-based cancer research, 2021.

[NST02] Nygård, J., Skare, G., and Thoresen, S. "The cervical cancer screening programme in Norway, 1992–2000: changes in Pap smear coverage and incidence of cervical cancer". In: *Journal of medical screening* vol. 9, no. 2 (2002), pp. 86–91.

[Ost93] Ostör, A. "Natural history of cervical intraepithelial neoplasia: a critical review." In: *International journal of gynecological pathology: official journal of the International Society of Gynecological Pathologists* vol. 12, no. 2 (1993), pp. 186–192.

[Par+15] Parmar, C. et al. "Machine learning methods for quantitative radiomic biomarkers". In: *Scientific reports* vol. 5, no. 1 (2015), pp. 1–11.

[Par+19] Partanen, V.-M. et al. "NordScreen–an interactive tool for presenting cervical cancer screening indicators in the Nordic countries". In: *Acta Oncologica* vol. 58, no. 9 (2019), pp. 1199–1204.

[Ped+17] Pedersen, K. et al. "Advancing the evaluation of cervical cancer screening: development and application of a longitudinal adherence metric". In: *The European Journal of Public Health* vol. 27, no. 6 (2017), pp. 1089–1094.

[Ped+18] Pedersen, K. et al. "An overview of cervical cancer epidemiology and prevention in Scandinavia". In: *Acta obstetricia et gynecologica Scandinavica* vol. 97, no. 7 (2018), pp. 795–807.

[Ras15] Raschka, S. *Python machine learning*. Packt publishing ltd, 2015.

[RHW85] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[Rod+08] Rodriguez, A. C. et al. "Rapid clearance of human papillomavirus and implications for clinical focus on persistent infections". In: *Journal of the National Cancer Institute* vol. 100, no. 7 (2008), pp. 513–517.

[Rot+18] Rothberg, M. B. et al. "A risk prediction model to allow personalized screening for cervical cancer". In: *Cancer Causes & Control* vol. 29, no. 3 (2018), pp. 297–304.

[Sam+21] Sambasivan, N. et al. ""Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI". In: *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–15.

[Sar21] Sarker, I. H. "Machine learning: Algorithms, real-world applications and research directions". In: *SN Computer Science* vol. 2, no. 3 (2021), pp. 1–21.

[Sch+07a] Schiffman, M. et al. "A 2-year prospective study of human papillomavirus persistence among women with a cytological diagnosis of atypical squamous cells of undetermined significance or low-grade squamous intraepithelial lesion". In: *The Journal of infectious diseases* vol. 195, no. 11 (2007), pp. 1582–1589.

[Sch+07b] Schiffman, M. et al. "Human papillomavirus and cervical cancer". In: *The lancet* vol. 370, no. 9590 (2007), pp. 890–907.

[Sch+16a] Schiffman, M. et al. "Carcinogenic human papillomavirus infection". In: *Nature reviews Disease primers* vol. 2, no. 1 (2016), pp. 1–20.

[Sch+16b] Schnabel, T. et al. "Recommendations as treatments: Debiasing learning and evaluation". In: *international conference on machine learning*. PMLR. 2016, pp. 1670–1679.

[SJ03] Srebro, N. and Jaakkola, T. "Weighted low-rank approximations". In: *Proceedings of the 20th international conference on machine learning (ICML-03)*. 2003, pp. 720–727.

[Son+21] Soni, A. et al. "Design of a machine learning-based self-driving car". In: *Machine Learning for Robotics Applications*. Springer, 2021, pp. 139–151.

[Sop+20] Soper, B. C. et al. "A hidden Markov model for population-level cervical cancer screening data". In: *Statistics in Medicine* vol. 39, no. 25 (2020), pp. 3569–3590.

[Sun+21] Sung, H. et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* vol. 71, no. 3 (2021), pp. 209–249.

[SW13a] Schiffman, M. and Wentzensen, N. "Human papillomavirus infection and the multistage carcinogenesis of cervical cancer". In: *Cancer Epidemiology and Prevention Biomarkers* vol. 22, no. 4 (2013), pp. 553–560.

[SW13b] Schiffman, M. and Wentzensen, N. "Human papillomavirus infection and the multistage carcinogenesis of cervical cancer". In: *Cancer epidemiology, biomarkers & prevention* vol. 22, no. 4 (2013), pp. 553–560.

[Tak+08] Takács, G. et al. “Matrix factorization and neighbor based algorithms for the netflix prize problem”. In: *Proceedings of the 2008 ACM conference on Recommender systems*. 2008, pp. 267–274.

[Tay+21] Tayefi, M. et al. “Challenges and opportunities beyond structured data in analysis of electronic health records”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* vol. 13, no. 6 (2021), e1549.

[Vac+14] Vaccarella, S. et al. “50 years of screening in the Nordic countries: quantifying the effects on cervical cancer incidence”. In: *British journal of cancer* vol. 111, no. 5 (2014), pp. 965–969.

[VEP21] Vuttipittayamongkol, P., Elyan, E., and Petrovski, A. “On the class overlap problem in imbalanced data classification”. In: *Knowledge-based systems* vol. 212 (2021), p. 106631.

[Wal+99] Walboomers, J. M. et al. “Human papillomavirus is a necessary cause of invasive cervical cancer worldwide”. In: *The Journal of pathology* vol. 189, no. 1 (1999), pp. 12–19.

[WM97] Wolpert, D. H. and Macready, W. G. “No free lunch theorems for optimization”. In: *IEEE transactions on evolutionary computation* vol. 1, no. 1 (1997), pp. 67–82.

[Yan+19] Yang, W. et al. “Cervical cancer risk prediction model and analysis of risk factors based on machine learning”. In: *Proceedings of the 2019 11th International Conference on Bioinformatics and Biomedical Technology*. 2019, pp. 50–54.

[YRD16] Yu, H.-F., Rao, N., and Dhillon, I. S. “Temporal regularized matrix factorization for high-dimensional time series prediction”. In: *Advances in neural information processing systems*. 2016, pp. 847–855.

[Zha+20] Zhang, S. et al. “Cervical cancer: Epidemiology, risk factors and screening”. In: *Chinese Journal of Cancer Research* vol. 32, no. 6 (2020), p. 720.

[Zho+11] Zhou, J. et al. “A multi-task learning formulation for predicting disease progression”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011, pp. 814–822.

[Zho+14] Zhou, J. et al. “From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2014, pp. 135–144.

[AAO17] Awoyemi, J. O., Adetunmbi, A. O., and Oluwadare, S. A. “Credit card fraud detection using machine learning techniques: A comparative analysis”. In: *2017 international conference on computing networking and informatics (ICCNI)*. IEEE. 2017, pp. 1–9.

Papers

Paper I

Matrix factorization for the reconstruction of cervical cancer screening histories and prediction of future screening results

Geir Severin R E Langberg, Mikal Stapnes, Jan F Nygård, Mari Nygård, Markus Grasmair, Valeriya Naumova

Published in *BMC Bioinformatics Supplements*, 2022, DOI: 10.1000/182.

METHODOLOGY

Matrix factorization for the reconstruction of cervical cancer screening histories and prediction of future screening results

Geir Severin R E Langberg[1*], Mikal Stapnes[3], Jan F. Nygård[2], Mari Nygård[1], Markus Grasmair[3] and Valeriya Naumova[4]

[*]Correspondence: severin.langberg@kreftregisteret.no
[1]Department of Research, Cancer Registry of Norway, Ullernchausseen 64, 0379 Oslo, Norway
Full list of author information is available at the end of the article

Abstract

Background:

Mass screening programs for cervical cancer prevention in the Nordic countries have strongly reduced cancer incidence and mortality at the population level. An alternative to the current mass screening is a more personalised screening strategy adapting the recommendations to each individual. However, this necessitates reliable risk prediction models accounting for disease dynamics and individual data.

Herein we propose a novel matrix factorisation framework to classify females by the time-varying risk of being diagnosed with cervical cancer. We cast the problem as a time-series prediction model where the data from females in the Norwegian screening population are represented as sparse vectors in time and then combined into a single matrix. Using novel temporal regularisation and discrepancy terms for the cervical cancer screening context, we reconstruct complete screening profiles from this scarce matrix and use these to predict the next exam results indicating the risk of cervical cancer. The algorithm is validated on both synthetic and registry screening data by measuring the *probability of agreement* (PoA) between Kaplan-Meier estimates.

Results:

In numerical experiments on synthetic data, we demonstrate that the novel regularisation and discrepancy term can improve the data reconstruction ability as well as prediction performance over varying data scarcity. Using a hold-out set of screening data, we compare several numerical models and find that the proposed framework attains the strongest PoA. We observe strong correlations between the empirical survival curves from our method and the hold-out data, and evaluate the ability of our framework to predict the females' next results for up to five years ahead in time using only their current screening histories as input.

Conclusions:

We have proposed a matrix factorization model for predicting future screening results and evaluated its performance in a female cohort to demonstrate the potential for developing prediction models for more personalized cervical cancer screening.

Keywords: cervical cancer; cancer screening; population-level cancer prevention; matrix completion; matrix factorization

Background

The mass screening programs against cervical cancer established in the Nordic countries may have prevented up to 80% of malignancies [1]. Persistent Human papillomavirus (HPV) infection is the primary causes of cervical cancer – as well as several other cancer types – initiating a process of cellular changes from low-grade to high-grade (pre-cancerous) lesions to invasive cancer [2]. Early detection of pre-cancerous lesions, e.g. with cytology, histology, or HPV tests, could prevent cancer development if it is treated [3] and motivates the need for screening.

A key factor in the success of the cancer screening programs is repeated screening at regular intervals. However, the risk of being infected with HPV and the risk of progressing to cancer vary significantly between females [4]. Thus, too frequent screening may lead to over-treatment of clinically insignificant pre-cancers, while too infrequent screening risks missing pre-cancers warranting treatment.

An alternative to the current mass-screening is a more personalized strategy adapting the screening frequency to the individual risk of disease initiation. For instance, vaccination of adolescent females has shown to improve protection against HPV infection [5], in which case the cancer screening programs may benefit from more flexible guidelines for the individual risk [6]. A step towards guidelines for more personalized recommendations is developing prediction models for the time-varying risk of cervical cancer using existing screening data from centrally organized population-level registries. In this paper, we present a novel matrix factorisation framework for time-dependent risk assessment of cervical cancer. We use population-based data from the *Norwegian Cervical Cancer Screening Program* (NCCSP) and evaluate our method by comparing Kaplan-Meier estimators from model predictions and a hold-out set.

The NCCSP database contains only the information needed by the *Cancer Registry of Norway* to administer the screening program. There are test results from 3 types of medical exams (cytology, histology, and HPV) but no further clinical information about the NCCSP participants. Following [7] we process these results into four *states*, reflecting the risk of cervical cancer and clinical consequences: A *normal* state indicates an accepted baseline risk; a *low-risk* state indicating an early stage of carcinogenesis (low-grade lesion) warranting more frequent screening to catch a potential progression to *high-risk*, requiring immediate treatment, and a *cancer* state, which can be seen as a failure of the screening program and a potentially lethal state for the woman.

In our approach we use NCCSP data collected between 1991–2015. During this time period, females aged 25–69 with a prior normal result were invited to a routine screening every 3rd year. According to those guidelines, triennial screening amounts to about 15 results in total and thus the state of the cervix is only observed at a few time points (*scarce* data). Moreover, since the recommendations are not strictly adhered to in practice the individual screening histories become *irregular* over time. Lastly, the majority of exam results are normal, making the data highly imbalanced. Specifically, in the NCCSP more than 90 % of test results are normal, 4–5 % low-risk and around 1 % are high-risk or cancer [8].

In Figure 1, we illustrate screening histories represented by sparse time series vectors fitted into a matrix. Our goal is to estimate complete state profiles by filling

the missing entries of these vectors and then use the completed state profiles in predicting the future state. Assuming correlation between subgroups of screening histories, we estimate the complete profiles using low-rank matrix factorisation (MF) and matrix completion (MC) techniques.

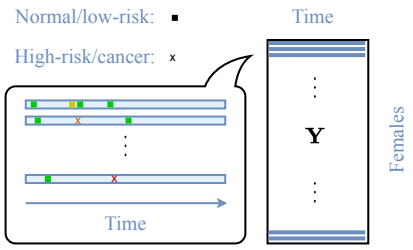

Figure 1 Matrix representation of cervical cancer screening histories. Individual cervical cancer screening histories as sparse time series fitted into a matrix $\mathbf{Y}$. A green/yellow square indicates normal/low-risk state, and an orange cross denotes a high-risk/cancer state. The matrix columns corresponds to female age intervals of 3 months.

Existing methods applying MF to temporal data use similarity networks encoding temporal dependencies to facilitate constraints on the solution [9]. However, in our case the explicit temporal structure is not easily inferred from the data. Some recent work [10] extends the geometric deep learning (GDL) framework [11] to the matrix completion problem. Similarly to the temporal MF approaches, geometric deep learning methods also encode the structure of the data matrix using similarity graphs. The *PACIFIER* framework is a MF approach [12] specifically targeting the healthcare domain and the analysis of Electronic Medical Records, which can also be very sparse and noisy similar to the screening data. The PACIFIER performs MC by imposing sparsity and smoothness constraints on the temporal evolution of the latent factors.

In this paper, we adapt the PACIFIER framework to the cervical cancer screening setting and reconstruct complete state profiles from the scarce histories. We present a regulariser for the temporal dependencies between the results in histories and propose a discrepancy term for utilizing correlations between different histories. We evaluate our method on both synthetic data and registry data by measuring the *probability of agreement* [13] between Kaplan-Meier estimates from model predictions and a hold-out set.

Results

In our experiments we consider five matrix factorization methods. The first method, referred to only as *matrix factorization* (MF), is our implementation of the PACIFIER. The second method, *convolutional* MF (CMF), extends the PACIFIER with more flexibility to model the variability observed in the cancer screening data. Furthermore, we introduce time shifts into the CMF to better exploit correlations

between screening histories and name this *shifted* CMF (SCMF). We also consider versions of the CMF and SCMF where the errors in the discrepancy term are weighted to emphasize particular exam results. These models are referred to in our experiments as *weighted* CMF (WCMF) and *weighted* SCMF (WSCMF).

Moreover, we compare the matrix factorization models to the *GDL approach for matrix completion* (GDL) as in [10]. We studied different ways of constructing similarity graphs capturing the structure on the rows and columns of our matrix representation of screening histories, $\mathbf{Y}$, as input to GDL. Our strongest results over various distance metrics, including Euclidean and Wasserstein distance, came with a 10-NN sequential column graph for temporal smoothness and a 10-NN row graph based on the *cosine distance* to connect similar screening histories. Both graphs are weighted by $\exp(-d(i,j))$ with $d(i,j)$ being the distance between two connected nodes i and j.

Synthetic data experiments

We generated synthetic data resembling the scarcity, irregularity and imbalance of the registry screening data. Latent state profiles were synthesized from linear combinations of five basic profiles of the form $V_{t,k} = \exp(-10^{-3}(t-\mu_k)^2)$ and female-specific coefficients $U_{n,k} \sim \text{Exp}(1)$. We mapped each of the entries in the latent state matrix $\mathbf{M} \in \mathbb{R}^{N\times T}$ to an integer 1–4 with model (2) at $\theta = 2.5$. Entries were randomly removed from the resulting integer matrix using empirical probabilities of observing an entry conditioned on the previous state. Figure 2 compares the synthetic data and the cancer screening registry data.

To measure the reconstruction error between the model estimate $\widehat{\mathbf{M}}$ and the ground truth $\mathbf{M}$ over the unobserved entries, we use

$$\mathcal{D} \triangleq \frac{\left\| \mathcal{P}_{\Omega^c}\left(\mathbf{M} - \widehat{\mathbf{M}}\right)\right\|_F^2}{N\,T\,\overline{|\Omega^c|}}. \tag{1}$$

The operator $\mathcal{P}_{\Omega^c}\colon \mathbb{R}^{N\times T} \to \mathbb{R}^{N\times T}$ projects onto unobserved entries and $\overline{|\Omega^c|}$ is the fraction of entries from $\mathbf{Y}$ in Ω^c. Figure 3 shows the reconstruction error for factorization models MF, CMF and SCMF over varying data density $\overline{|\Omega|}$ given as the fraction of observed entries.

Figure 3 indicates that the temporal regularisation used in CMF produces more accurate data reconstructions than the regularisation used in MF as reconstruction error is consistently smaller for CMF than for MF. Moreover, the shift mechanism in SCMF, exploiting correlations between screening histories, gives even smaller reconstruction errors.

In Figure 4 we compare performance scores, Φ_s (Eq. (8)) for different models, indicating the *probability of agreement* [13] between hold-out data and predictions. Predicting based on Eq. (5), we required at least two results to be observed prior to the prediction time and in addition we used a moving window to ensure that no result was observed within two years from the time to predict.

The PoA-based scores in Figure 4 shows that SCMF typically achieves the strongest performance, followed by CMF, mostly outperforming MF. Especially in classifying normal and low-risk, where the number of cases is higher than for high-risk and cancer, the SCMF and CMF attain the highest scores.

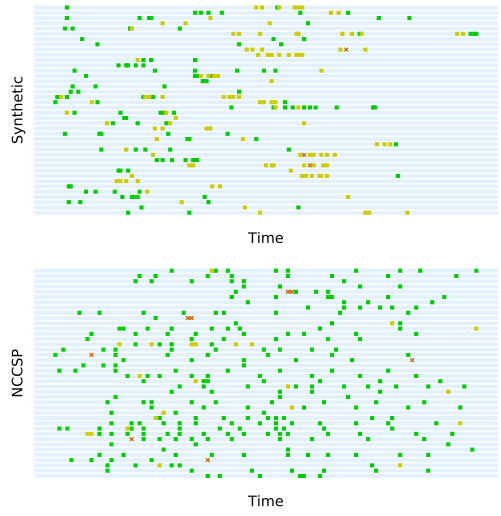

Figure 2 Comparing synthetic data to screening data. Randomly selected histories from synthetic data and data from the Norwegian Cervical Cancer Screening Programme (NCCSP). Green/yellow squares correspond to normal/low-risk results, and an orange cross signify either high-risk or cancer results.

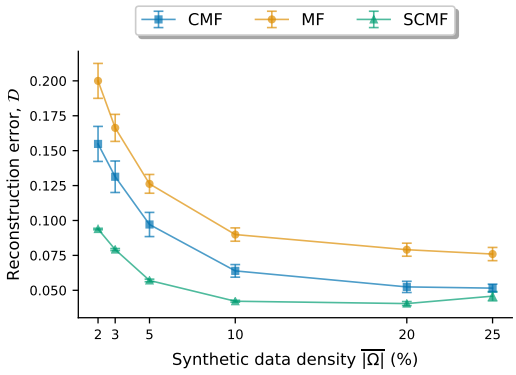

Figure 3 Reconstruction error on synthetic data. Comparing the reconstruction error ($\mathcal{D}$ by (1)) from different factorization models specified in Table 3 over varying data density ($\overline{|\Omega|}$). The factorization models are Matrix Factorization (MF), Convolutional MF (CMF) and Shifted CMF (SCMF).

Screening data experiments

We randomly sampled two sets of 15K screening histories (training and test) with at least 3 results between 1991–2015 from the NCCSP data including over 1.7 million female participants. Each selected female was born between 1965–1970 and had

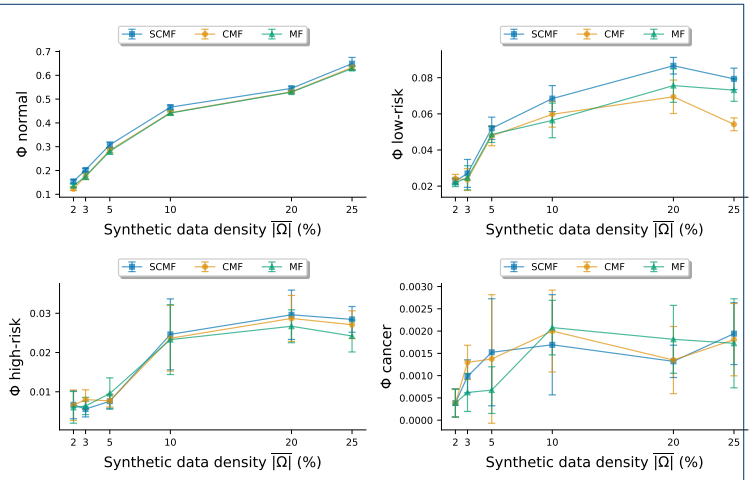

Figure 4 Probability of agreement on synthetic data. Classification performance on synthetic data of varying data density $\overline{|\Omega|}$. Model performance is given as the *probability of agreement* [13] score (Φ_s from (8)) with 95% CI. Higher $\Phi_s \in [0, 1]$ signifies better model fit. The prediction models are Matrix Factorization (MF), Convolutional MF (CMF) and Shifted CMF (SCMF).

her first screening at age 25 (the recommended age to start screening by NCCSP guidelines) to minimize left-censoring. Organizing the histories as sparse time series and combining them produced training and test matrices, each with about 8% observed entries.

The training histories were used to estimate latent state profiles with the models from Table 3 and a GDL based on [10]. Classification thresholds were obtained by solving (6). The test histories were used for model performance evaluation by comparing observed and predicted results over time, like in experiments on synthetic data. Table 1 gives the normalized PoA score (Φ_s; Eq. (8)) per prediction model.

Table 1 Classification performance on registry screening data. Model performance is given as the *probability of agreement* [13] score (Φ_s) with 95% CI. Higher $\Phi_s \in [0, 1]$ signifies better model fit.

	Φ_s				
Model	Normal	Low-risk	High-risk	Cancer	$\sum \Phi_s$
GDL	0.35 [0.32, 0.43]	0.087 [0.077, 0.094]	0.15 [0.13, 0.17]	0.47 [0.44, 0.51]	1.1
MF	0.28 [0.22, 0.35]	0.022 [0.00, 0.063]	0.21 [0.19, 0.24]	0.46 [0.33, 0.54]	0.98
CMF	0.31 [0.23, 0.39]	0.11 [0.063, 0.12]	0.29 [0.27, 0.32]	0.77 [0.72, 0.83]	1.5
WCMF	0.31 [0.26, 0.35]	0.25 [0.23, 0.27]	0.27 [0.24, 0.31]	0.78 [0.73, 0.87]	1.6
SCMF	0.33 [0.27, 0.39]	**0.59 [0.57, 0.62]**	**0.35 [0.32, 0.37]**	0.63 [0.55, 0.71]	1.9
SWCMF	**0.36 [0.29, 0.41]**	0.50 [0.47, 0.51]	0.33 [0.24, 0.41]	**0.86 [0.80, 0.90]**	**2.1**

The overall PoA score in Table 1 was highest for SWCMF from being the most accurate model to predict normal ($\Phi_s = 0.36$) and cancer ($\Phi_s = 0.86$). High-risk and low-risk was best predicted by SCMF ($\Phi_s = 0.35$ and $\Phi_s = 0.59$). Note that CMF improves on MF and both shifted models (SWCMF and SCMF) mostly outperformed their non-shifted variants.

Based on achieving the highest overall PoA score, we study SWCMF in classifying with a forecast horizon ranging from 0.5–5 years. The SWCMF performances from

predicting with all data within a given time from the target being removed are given in Table 2.

Table 2 Classification performance for Shifted Weighted Convolutional Matrix Factorization over varying forecast horizon as the *probability of agreement* [13] score (Φ_s from 8) with 95% CI. Higher $\Phi_s \in [0, 1]$ signifies better model fit.

	Φ_s				
Forecast (years)	Normal	Low-risk	High-risk	Cancer	$\sum \Phi_s$
0.5	0.35 [0.26, 0.40]	0.61 [0.52, 0.63]	0.21 [0.18, 0.24]	0.91 [0.86, 0.95]	2.1
1	0.32 [0.25, 0.36]	0.59 [0.56, 0.62]	0.45 [0.35, 0.52]	0.90 [0.83, 0.96]	2.3
2	0.36 [0.29, 0.41]	0.50 [0.47, 0.51]	0.33 [0.24, 0.41]	0.86 [0.80, 0.90]	2.1
3	0.38 [0.33, 0.43]	0.40 [0.38, 0.41]	0.24 [0.21, 0.26]	0.79 [0.70, 0.85]	1.8
5	0.20 [0.086, 0.29]	0.024 [0.020, 0.025]	0.20 [0.10, 0.28]	0.68 [0.66, 0.73]	1.1

Table 2 shows that the SWCMF performance is relatively stable up to 3 year forecasts, which is the longest recommended exam interval. However, the performance drops noticeably at the 5 year forecast.

Plotting the Kaplan-Meier estimates for the hold-out set and the 2 year SWCMF predictions in Figure 5 indicates a good overall fit as model predictions clearly correlate with the observed data. Note that the y-axis scale differs between the plots.

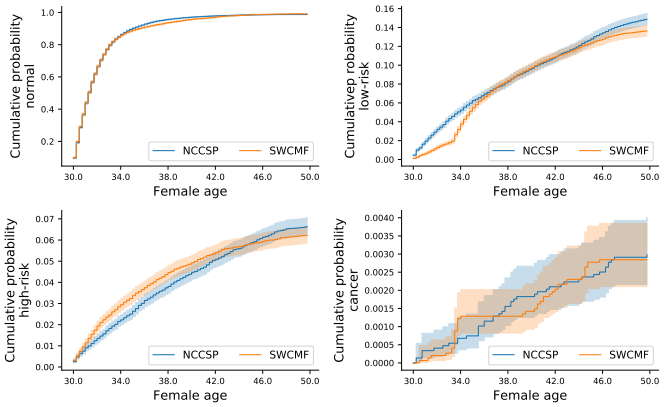

Figure 5 Kaplan-Meier estimates. Comparing Kaplan-Meier estimates from 2 year predictions with Shifted Weighted Convolutional Matrix Factorization (SWCMF) and a hold-out set of registry data from the Norwegian Cervical Cancer Screening Programme (NCCSP).

In Figure 5, the normal rate is slightly underestimated over ages 34–42, as well as the low-risk rate for younger (ages 30–36) and older (ages 44–50) females. These 3 regions correspond well to the times when high-risk is overestimated, which is likely the result of our method for setting the probability thresholds by solving (6). Using time-varying probability thresholds could potentially improve the results here.

The PoA curves from Kaplan-Meier estimates in Figure 5 are plotted in Figure 6 to evaluate their agreement.

According to Figure 6 there is a strong agreement between the cancer estimates, especially after around age 40. As observed in Figure 5, the drop in PoA for high-risk

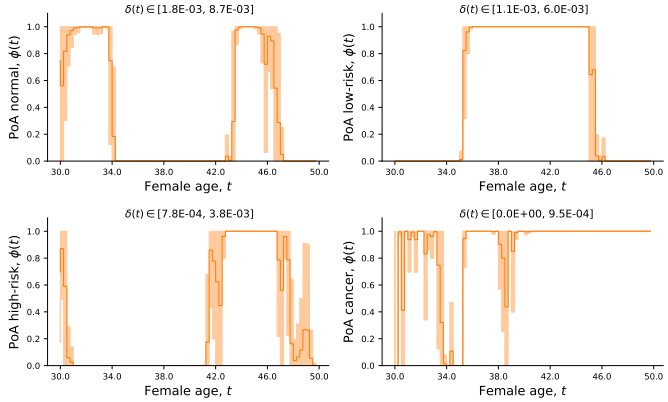

Figure 6 Probability of agreement. The probability of agreement ((PoA); $\phi(t)$ from (7)) between the Kaplan-Meier estimates in Figure 5. Higher $\phi(t) \in [0, 1]$ means stronger agreement. The range of equivalence margins is given as $\delta(t) \in [a, b]$.

is complementary to the PoA for normal and low-risk, in which case overestimating high-risk leads to underestimating low-risk and normal in our classification model.

Discussion and conclusions

Deriving risk prediction models from existing cancer screening registries is a step towards more personalized screening. Here we present a matrix factorization framework that, to our knowledge, is the first approach to use this method for classifying females by the time-varying risk of being diagnosed with cervical cancer from only their current screening histories.

Here we used screening histories from females participating in the Norwegian Cervical Cancer Screening Programme (NCCSP) between 1991–2015, and represent these as sparse time-series vectors fitted into a single matrix. Comparing different algorithms for estimating complete screening profiles for each female we found that the proposed framework, accounting for temporal dependencies within histories and correlations between samples, gave the most accurate estimates.

To illustrate the potential for developing risk prediction models for more personalized screening recommendations, we validated the framework on the NCCSP registry data using Kaplan-Meier (K-M) estimates from model predictions and a hold-out set. The K-M curves showed a strong correlation and a corresponding high *probability of agreement* (PoA) [13] using an equivalence margin $(-\delta(t), \delta(t))$ based the time-varying standard deviation of the ground truth K-M curve.

A typical choice to check if two quantities are within $q\,\%$ of each other is $\delta = q/100$, but this fixed margin does not permit potential temporal variation in the similarity measure depending on the uncertainty in the reference data. Using the time-varying standard deviation for margin, as in our case, gives a more strict measure if the uncertainty in the ground truth K-M estimate is small but may potentially increase the PoA if this estimate has high variance As the choice for δ greatly affects the PoA

measure, methods for selecting this parameter in cervical cancer screening contexts should be addressed in future work.

Adapting screening recommendations to females at reduced or elevated risk may improve efficiency and precision of cancer screening programs. Prediction models for the individual risk can assist screening programs in adapting to such personalized strategies. The framework presented herein demonstrates the potential for using matrix factorization to derive prediction models for personalized risk estimation based on individual screening data. We also believe that our approach could be applied to data from other types of mass-screening programmes such as breast, colorectal and prostate cancer, which we plan to investigate in future work.

Methods

We represent the cervical cancer screening data as a partially observed matrix $\mathbf{Y} \in \mathbb{N}^{N\times T}$. Each row in $\mathbf{Y}$ is a one-dimensional time series for a single screening history and each column represents a 3 months time interval. Based on recommendations of 3 years screening intervals for healthy females, and 3 to 6 months for females at elevated risk, choosing 3 months for the time discretisation of the data provides thus a reasonable compromise between temporal resolution and sparsity of the data. In the following, we denote the set of indices where observations in $\mathbf{Y}$ are available by $\Omega \subset \{n\}_{n=1}^{N} \times \{t\}_{t=1}^{T}$. Moreover, each observed entry $Y_{n,t} \in \mathbf{Y}$, representing a *normal*, *low-risk*, *high-risk* or a *cancer* state, is numerically encoded with integer values $s \in \{1, 2, 3, 4\}$ where 1 is normal and 4 is cancer, as in [7].

A latent state model for cervical cancer screening data

Our basic assumption is that the discrete observed states $Y_{n,t}$ are possibly inaccurate measurements of a continuous *latent state* $M_{n,t}$ that evolves slowly over time for each female. We take each state $Y_{n,t}$ to be observed with probability based on a Gaussian distribution of mean $M_{n,t}$ and variance $1/2\theta$. The parameter $\theta > 0$ models the reliability of the estimate. Thus,

$$p(Y_{n,t} = s \mid M_{n,t}) \triangleq C_{M_{n,t}} \exp(-\theta(s - M_{n,t})^2) \tag{2}$$

for some normalization constant $C_{M_{n,t}}$. With this model we have the maximum likelihood estimate

$$\theta^\star = \frac{|\Omega|}{2\sum_{(n,t)\in\Omega} (Y_{n,t} - M_{n,t})^2},$$

where $|\Omega|$ is the number of observations in Ω.

Furthermore, we assume that each latent state profile is a linear combination of a small number of basic profiles $\mathbf{v}_1, \ldots, \mathbf{v}_r$ with $r \ll \min\{N, T\}$. Then the matrix $\mathbf{M}$ of all such profiles can be approximately decomposed as $\mathbf{M} \approx \mathbf{U}\mathbf{V}^\top$ with $\mathbf{V} \in \mathbb{R}^{T\times r}$ being the collection of basic profiles and $\mathbf{U} \in \mathbb{R}^{N\times r}$ being the female-specific coefficients. Figure 7 illustrates the latent state model.

For the simultaneous reconstruction of $\mathbf{U}$ and $\mathbf{V}$, we propose the variational method of solving

$$\min_{\mathbf{U},\mathbf{V}} \left\{ \left\| \mathbf{W} \odot \left(\mathbf{Y} - \mathbf{U}\mathbf{V}^\top\right) \right\|_F^2 + \alpha_1 \|\mathbf{U}\|_F^2 + \alpha_2 \|\mathbf{V}\|_F^2 + \alpha_3 \|\mathbf{R}\mathbf{V}\|_F^2 \right\}. \tag{3}$$

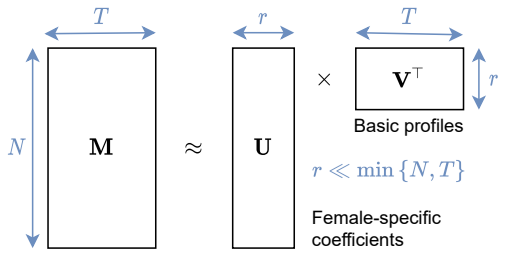

Figure 7 Latent state model for cervical cancer screening data. The matrix $\mathbf{M} \in \mathbb{R}^{N\times T}$ of latent state profiles decomposed into female-specific coefficients $\mathbf{U} \in \mathbb{R}^{N\times r}$ and a small number ($r \ll \min\{N, T\}$) of basic profiles $\mathbf{V} \in \mathbb{R}^{T\times r}$.

Here, $\mathbf{W} \in \mathbb{R}^{N\times T}$ sets all matrix entries $(\tilde{n}, \tilde{t}) \notin \Omega$ to 0 and multiplies the error over the predicted values at the observed entries $(n, t) \in \Omega$ with some weights $W_{n,t} > 0$. These weights provide a flexible way to incorporate additional information such as uncertainties in exam results and adjusting for entries $Y_{n,t}$ not missing at random with inverse propensity weighting [14]. The matrix $\mathbf{R} \in \mathbb{R}^{N\times N}$ is used to enforce some time-regularity on the basic profiles $\mathbf{v}_1, \cdots, \mathbf{v}_r$. We consider two choices of $\mathbf{R}$, the first being the forward difference matrix $\mathbf{R} = \mathbf{D}$. This has the effect of enforcing a high temporal smoothness and is in line with the approach of [12]. As an alternative, we propose $\mathbf{R} = \mathbf{KD}$ with the forward difference matrix $\mathbf{D}$ and $\mathbf{K}$ being the Toeplitz matrix with entries $K_{ij} = \exp(-\gamma\,|i - j|)$. This leads to a weaker penalisation of the profiles at faster scales and consequently allows for a larger local variability. The same variability is also observed in the NCCSP data as long intervals with normal results followed by rapid recurrent exams after an abnormal result is detected.

In the NCCSP data we also observe strong correlations between screening histories although as slightly shifted in time. To better exploit these correlations, we extend (3) with female-specific shift matrices $\mathbf{Z}_n \in \{0, 1\}^{T\times T}$ containing ones in the z_n-th diagonal and zeros everywhere else. Now $\mathbf{V}^\top\mathbf{Z}_n$ shifts the basic profiles $z_n \in \mathbb{Z}$ time points either forward ($z_n > 0$) or backward ($z_n < 0$) to improve alignment with screening history $\mathbf{Y}_n$. We limit z_n to at most 3 years shift forward or backward in time. To simultaneously optimize $\mathbf{U}$, $\mathbf{V}$ and the vector $\mathbf{z}$ of N offset values, we solve

$$\min_{\mathbf{U},\mathbf{V},\mathbf{z}} \left\{ \sum_{n=1}^{N} \left\| \mathbf{W}_n \odot \left(\mathbf{Y}_n - \mathbf{U}_n\mathbf{V}^\top\mathbf{Z}_n\right)\right\|_F^2 + \beta_1 \sum_{n=1}^{N} \|\mathbf{U}_n\|_2^2 + \beta_2 \|\mathbf{V}\|_F^2 + \beta_3 \|\mathbf{RV}\|_F^2 \right\}. \tag{4}$$

Here $\mathbf{W}_n$, $\mathbf{Y}_n$ and $\mathbf{U}_n$ are vectors from the n-th row of each matrix.

Following [12], we optimize (3) by alternating between solving for $\mathbf{U}$ at fixed $\mathbf{V}$, and solving for $\mathbf{V}$ at fixed $\mathbf{U}$. To optimize $\mathbf{z}$ in (4) we add an exhaustive search over candidate z_n. In numerical experiments, we initialize the iterations with $V_{t,k} \sim \mathcal{N}(0, 1)$ and $\mathbf{z}$ as a vector of zeros. The iterations abort once the relative difference between consecutive estimates $\widehat{\mathbf{M}}^{(i+1)}$ and $\widehat{\mathbf{M}}^{(i)}$ is less than 10^{-6}.

Based on the models (3) and (4), we define five factorization models used in numerical experiments. Table 3 characterizes the factorization models by temporal smoothness model $\mathbf{R}$, discrepancy weights $W_{n,t} \in \mathbf{W}$ and female-specific shifts z_n.

Table 3 Matrix factorization models ((3) and (4)) used in numerical experiments.

Model name	$\mathbf{R}$	$W_{n,t} : (n,t) \in \Omega$	$\max z_n$ (years)
Matrix Factorization (MF)	$\mathbf{D}$	1	-
Convolutional MF (CMF)	$\mathbf{KD}$	1	-
Shifted CMF (SCMF)	$\mathbf{KD}$	1	3
Weighted CMF (WCMF)	$\mathbf{KD}$	$\widehat{p}(s \mid \epsilon) \,/\, \widehat{p}((n,t) \in \Omega)$	-
Shifted WCMF (SWCMF)	$\mathbf{KD}$	$\widehat{p}(s \mid \epsilon) \,/\, \widehat{p}((n,t) \in \Omega)$	3

As specified in Table 3, the weights in WCMF and SWCMF incorporate inverse propensity weighting. for our experiments, we derived propensity estimates $\widehat{p}((n,t) \in \Omega)$ using the method in [15] and uncertainties in the medical the exam types (i.e., cytology or histology) from [16].

Predicting the next screening result

To evaluate the proposed framework, we compare here Kaplan-Meier estimates from model predictions with a hold-out set. In future work we plan to evaluate our method for the prediction of individual results.

To predict the future state of a single female, we assume that we are given her current screening record $\mathbf{x} \in \mathbb{N}^T$ with observations at times $t_0 \leq t_p, \ldots, t_q < T$, and that $\mathbf{m} \in \mathbf{M}$ is the latent state profile underlying $\mathbf{x}$. To predict a future state s at $t_{q+1} > t_q$, we consider the conditional probability

$$p(x_{t_{q+1}} = s \mid \mathbf{x}) \propto \int p(x_{t_{q+1}} = s \mid \mathbf{m})\, p(\mathbf{m} \mid \mathbf{x})\, d\mathbf{m}.$$

Here $p(x_{t_{q+1}} = s \mid \mathbf{m})$ corresponds to model (2) and $p(\mathbf{m} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \mathbf{m})\pi(\mathbf{m})$ requires a prior $\pi(\mathbf{m})$ for profile $\mathbf{m}$. In our approach, we use the samples in $\widehat{\mathbf{M}}$ as a proxy for $p(\mathbf{m} \mid \mathbf{x})$. This yields the estimated conditional probabilities

$$\begin{aligned} \widehat{p}(x_{t_{q+1}} = s \mid \mathbf{x}) \propto & \sum_{n=1}^{N} C_{\widehat{M}_{n,t_{q+1}}} \exp(-\theta(s - \widehat{M}_{n,t_{q+1}})^2) \\ & \times \prod_{j=p}^{q} C_{\widehat{M}_{n,t_j}} \exp(-\theta(x_{t_j} - \widehat{M}_{n,t_j})^2). \end{aligned} \tag{5}$$

Applying estimator 5 to each value $s \in \{1,2,3,4\}$ gives a comprehensive probabilistic overview of a female's risk. To classify a female into a state from these risk estimates, we consider probability thresholds $\tau = \{\tau_s \in (0,1)\}_{s=2}^{4}$ as a way to alleviate the impact of data imbalance. Recall that in the registry data, the states are heavily skewed towards normal, which dominates the risk inference and bias predictions towards the normal state. For each state s, we check if the condition $\widehat{p}(x_{t_{q+1}} = s \mid \mathbf{x}) \geq \tau_s$ holds – in which case we predict $x_{t_{q+1}} = s$. The states are evaluated in order from $x_{t_{q+1}} = 4$ down to $x_{t_{q+1}} = 2$. This means that if the condition is satisfied for cancer ($s = 4$), we classify the female into a cancer state and

ignore the probabilities of high-risk and low-risk. If neither of the conditions are satisfied we predict normal ($x_{t_{q+1}} = 1$).

To select probability thresholds we first construct Kaplan-Meier estimates $\widehat{S}_s$ for each state from model predictions and the corresponding estimates S_s from the ground truths. An event in the Kaplan-Meier estimate is taken to be the first encounter of a specific state in the screening history of a female; if there are several events, we only record the first one. In the second step we solve

$$\min_{\tau} \sum_{s} \int_{t_0}^{T} |S_s(t) - \widehat{S}_s(t)|\, dt \tag{6}$$

to obtain the threshold values. Here we use the *differential evolution algorithm* [17] to search for threshold values although an exhaustive search could improve performance at the cost of higher computational complexity. The choice to minimize $|S_s(t) - \widehat{S}_s(t)|$ comes from our measure of model performance specified in the next section.

Model performance evaluation

As a way to assess the potential for developing prediction models for more personalized cervical cancer screening, we validate numerical models over a female cohort. We measure model performance as the *probability of agreement* (PoA) [13] between Kaplan-Meier estimates derived from model predictions and a holdout-set of screening data. This method relies on an appropriate choice of an indifference region $(-\delta, \delta)$ to determine the similarity between the two estimates.

At time $t \in [t_0, T]$ the PoA evaluates to

$$\phi_s(t) \triangleq p\big(|S_s(t) - \widehat{S}_s(t)| \leq \delta_s(t)\big). \tag{7}$$

Here $\phi_s(t)$ is the probability that the distribution of $S_s(t) - \widehat{S}(t)_s$ is contained within $\pm\delta$ to support a conclusion about the similarity of the true survival functions. A higher $\phi_s(t)$ implies that S_s and $\widehat{S}_s$ are more similar. Currently lacking scientific support for an indifference region eligible in cervical cancer screening, we simply let $\delta(t) = 2\,\widehat{\sigma}(S_s(t))$ estimated from 1000 bootstrap samples.

To quantify model performance in a single number, we estimate the normalized area under the PoA curve

$$\Phi_s \triangleq \frac{1}{T - t_0} \int_{t_0}^{T} \phi_s(t)\, dt. \tag{8}$$

Here $\Phi_s \in [0, 1]$ where $\Phi_s = 1$ indicates perfect model fit. We use the estimate in (8) to compare different models in numerical experiments.

List of abbreviations

- CI: Confidence interval
- CMF: Convolutional matrix factorization
- GDL: Geometric Deep Learning
- K-M: Kaplan-Meier

- MC: Matrix completion
- MF: Matrix factorization
- NCCSP: Norwegian Cervical Cancer screening Program
- NN: Nearest neighbour
- PoA: Probability of agreement
- SCMF Shifted convolutional matrix factorization
- SWCMF: Shifted weighted convolutional matrix factorization
- WCMF Weighted convolutional matrix factorization

Ethics approval and consent to participate
The project conducting this study is approved by the South East Norway Regional Committee for Medical and Health Research Ethics (application ID: 11752). All the research herein was performed in accordance with the relevant guidelines and regulations. The health registry data used in this study does not originate from clinical trials and therefore the ethical committee granted this study with an exception from informed consent.

Consent for publication
Not applicable.

Availability of data and materials
The cervical cancer screening datasets used in this study can be made available from the Cancer Registry of Norway pursuant the legal requirements mandated by the European GDPR, Article 6 and 9. The data are not publicly available due to individual privacy and ethical restrictions.
Source code (Python™) for synthetic data and numerical models can be provided by the corresponding author.

Competing interests
The authors declare that they have no competing interests.

Funding
This work is supported by the IKTPLUSS-program of the Research Council of Norway through the Decipher project (300034). The funder had no role in the design of the study, data collection, analysis or interpretation, or in writing the manuscript. Publication costs are covered by the Decipher project funds.

Author's contributions
GSREL, MS, VN and MG developed the model. GSREL and MS implemented the algorithms, and JFN contributed to framing the experiments. GSREL carried out the experiments. MN and JFN provided the registry cancer screening data and expertise on cervical cancer screening. All authors read and approved the final manuscript.

Acknowledgements
We thank Dr. Braden C. Soper (Lawrence Livermore National Laboratory) for useful discussions and advice.

Author details
[1]Department of Research, Cancer Registry of Norway, Ullernchausseen 64, 0379 Oslo, Norway. [2]Department of Registry Informatics, Cancer Registry of Norway, Ullernchausseen 64, 0379 Oslo, Norway. [3]Department of Mathematical Sciences, Norwegian University of Science and Technology, Høgskoleringen 1, 7491 Trondheim, Norway. [4]Department of Machine Intelligence, SimulaMet, Pilestredet 52, 0167 Oslo, Norway.

References
1. Vaccarella, S., Franceschi, S., Engholm, G., Lönnberg, S., Khan, S., Bray, F.: 50 years of screening in the Nordic countries: quantifying the effects on cervical cancer incidence. British journal of cancer **111**(5), 965–969 (2014)
2. Cohen, P.A., Jhingran, A., Oaknin, A., Denny, L.: Cervical cancer. Lancet **393**(10167), 169–182 (2019). doi:10.1016/S0140-6736(18)32470-X
3. WHO: Cervical Cancer. https://www.who.int/health-topics/cervical-cancer
4. Schiffman, M., Wentzensen, N.: Human papillomavirus infection and the multistage carcinogenesis of cervical cancer. Cancer Epidemiology and Prevention Biomarkers **22**(4), 553–560 (2013)
5. Laurent, J.S., Luckett, R., Feldman, S.: Hpv vaccination and the effects on rates of hpv-related cancers. Current problems in cancer **42**(5), 493–506 (2018)
6. Pedersen, K., Burger, E.A., Nygård, M., Kristiansen, I.S., Kim, J.J.: Adapting cervical cancer screening for women vaccinated against human papillomavirus infections: the value of stratifying guidelines. European Journal of Cancer **91**, 68–75 (2018)
7. Soper, B.C., Nygård, M., Abdulla, G., Meng, R., Nygård, J.F.: A hidden Markov model for population-level cervical cancer screening data. Statistics in Medicine (2020)
8. Nygård, J.F., Thoresen, S.O., Skare, G.B.: The cervical cancer screening program in Norway, 1992-2000. Changes in pap-smear coverage and cervical cancer incidence. International Journal of Cancer, 110–110 (2002)
9. Yu, H.-F., Rao, N., Dhillon, I.S.: Temporal regularized matrix factorization for high-dimensional time series prediction. In: Advances in Neural Information Processing Systems, pp. 847–855 (2016)
10. Monti, F., Bronstein, M.M., Bresson, X.: Geometric matrix completion with recurrent multi-graph neural networks. arXiv preprint arXiv:1704.06803 (2017)
11. Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P.: Geometric deep learning: going beyond Euclidean data. IEEE Signal Processing Magazine **34**(4), 18–42 (2017)

12. Zhou, J., Wang, F., Hu, J., Ye, J.: From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 135–144 (2014)
13. Stevens, N.T., Lu, L.: Comparing kaplan-meier curves with the probability of agreement. Statistics in Medicine **39**(30), 4621–4635 (2020)
14. Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., Joachims, T.: Recommendations as treatments: Debiasing learning and evaluation. In: International Conference on Machine Learning, pp. 1670–1679 (2016). PMLR
15. Ma, W., Chen, G.H.: Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. arXiv preprint arXiv:1910.12774 (2019)
16. Soper, B., Nygård, M., Abdulla, G., Meng, R., Nygård, J.F.: A Hidden Markov Model for population-level cervical cancer screening data. Technical report, Lawrence Livermore National Laboratory (2020)
17. Storn, R., Price, K.: Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization **11**(4), 341–359 (1997)

Paper II

Towards a data-driven system for personalized cervical cancer risk stratification

Geir Severin R E Langberg, Jan F Nygård, Vinay C Gogineni, Mari Nygård, Markus Grasmair, Valeriya Naumova

Published in *Scientific Reports*, 2022, DOI: 10.1038/s41598-022-16361-6.

II

scientific reports

OPEN

Towards a data-driven system for personalized cervical cancer risk stratification

Geir Severin R. E. Langberg[1]✉, Jan F. Nygård[2], Vinay Chakravarthi Gogineni[3], Mari Nygård[1], Markus Grasmair[4] & Valeriya Naumova[5]

Mass-screening programs for cervical cancer prevention in the Nordic countries have been effective in reducing cancer incidence and mortality at the population level. Women who have been regularly diagnosed with normal screening exams represent a sub-population with a low risk of disease and distinctive screening strategies which avoid over-screening while identifying those with high-grade lesions are needed to improve the existing one-size-fits-all approach. Machine learning methods for more personalized cervical cancer risk estimation may be of great utility to screening programs shifting to more targeted screening. However, deriving personalized risk prediction models is challenging as effective screening has made cervical cancer rare and the exam results are strongly skewed towards normal. Moreover, changes in female lifestyle and screening habits over time can cause a non-stationary data distribution. In this paper, we treat cervical cancer risk prediction as a longitudinal forecasting problem. We define risk estimators by extending existing frameworks developed on cervical cancer screening data to incremental learning for longitudinal risk predictions and compare these estimators to machine learning methods popular in biomedical applications. As input to the prediction models, we utilize all the available data from the individual screening histories. Using data from the Cancer Registry of Norway, we find in numerical experiments that the models are strongly biased towards normal results due to imbalanced data. To identify females at risk of cancer development, we adapt an imbalanced classification strategy to non-stationary data. Using this strategy, we estimate the absolute risk from longitudinal model predictions and a hold-out set of screening data. Comparing absolute risk curves demonstrate that prediction models can closely reflect the absolute risk observed in the hold-out set. Such models have great potential for improving cervical cancer risk stratification for more personalized screening recommendations.

Nation-wide cervical cancer screening programs in the Nordic countries have shown to be an effective cancer prevention strategy. These programs recommend repeated screening at regular intervals to detect precancerous lesions[1]. Although the screening recommendations have become more accurate and efficient over the years, they are based on only the most recent screening results and are standardised across the whole screening population. Specifically, the *Norwegian Cervical Cancer Screening Program* (NCCSP) currently recommends a routine screening every 3 or 5 years for females aged 25–33 years and 34–69 years, provided their last screening was normal. An alternative strategy would be to adapt the recommendations to the individual risk of disease initiation as inferred from the full screening history. For instance, a more personalized approach could be to recommend a longer screening interval to a female older than 45 who had only negative results in the past, as she may be at considerably lower risk than a 30 year old female with several past abnormalities. More personalized recommendations in cervical cancer screening may reduce the large number of unnecessary screenings of females unlikely to develop the disease, while simultaneously preventing more cancer cases[2].

A step towards more individualized recommendations is utilizing data from existing cancer screening registries to derive prediction models for the individual risk of cervical cancer development. However, the data available from these registries contain only a few variables about previous exam results, necessary to organize and run the screening programs but no information about female lifestyle or habits. Moreover, due to most

[1]Department of Research, Cancer Registry of Norway (CRN), Oslo 0379, Norway. [2]Department of Registry Informatics, CRN, Oslo 0379, Norway. [3]Department of Electronic Systems, Norwegian University of Science and Technology (NTNU), Trondheim 7491, Norway. [4]Department of Mathematical Sciences, NTNU, Trondheim 7491, Norway. [5]Machine Intelligence Department, Simula Research Laboratory, Oslo 0164, Norway. ✉email: langberg91@gmail.com

females having only normal results, the distribution of results is heavily skewed towards disease-free cases, and this distribution may also be changing over time due to temporal variations in female screening and lifestyle habits. In this paper, we use data from the NCCSP to evaluate the impact of data imbalance and data drift on model performance. We adapt machine learning methods to predict the individual time-varying risk of cervical cancer and compare their performances in numerical experiments.

Data. Our approach is based on data from the 1.7 million females in the NCCSP screening population between 1991–2015. As this data covers the Norwegian cervical cancer screening population, the prediction models derived herein can only be evaluated internally using hold-out methods. External model validation require data from a different country but differences in screening recommendations[3] and data collection practices make it challenging to align information for comparability. However, synergy projects with Baltic countries and Sweden are being developed to investigate the potential to extending predictors to other countries.

In this paper we considered only histories with at least 3 exam results for hold-out model validation. In the NCCSP data, more than 75 % of the females have only normal results in their history. To have more variation in the training data we sampled training histories with probabilities proportional to the most severe result in each history, making it more likely to select females with at least one abnormal result. For the test set we used only histories where the first exam was taken no later than year 2000 and at the ages 20–30 (± 5 years from the recommended youngest age for the first exam). This sampling give more recent and complete test data for a comprehensive model evaluation. As our dataset ends before 2015, selecting test histories with the first exam from year 2000 gave female age range 20–53 in the test data, while the results in the training set were from female ages 20–72. The final training set included 10K histories and the test set included 50K histories.

The NCCSP data contains only the information necessary for the *Cancer Registry of Norway* to organize and run the screening program. Although previous works[4,5] deriving prediction models for cervical cancer risk stratification leverage personal lifestyle information, this information is in general unavailable for the whole screening population. It is also typically collected only once for each female, and thus does not capture temporal variations in the data. Therefore, there is large potential and benefits in providing prediction models based on the data routinely collected by the registries, as these may be integrated directly into the cancer screening programs. Specifically, the NCCSP data consists of timestamps, three types of medical exams (*cytology*, *histology* and *human papillomavirus* (HPV)) and the corresponding results. The HPV exams were introduced around 2005 to follow up on abnormal cytology, and as our dataset ends before 2015 it contains only a few HPV results. Due to the scarcity of HPV results in our data sample, we exclude all HPV data in this study and use only cytology and histology results. However, we plan to include more recent registry data with detailed HPV information in future work.

The primary cause for cervical cancer is persistent infection with HPV. This infection may lead to the development of low-grade lesions, progressing via high-grade precursor lesions (pre-cancer) to invasive cancer[6,7]. Exposure to HPV occurs mainly via sexual contact which, together with individual lifestyle variations, makes the risk of cervical cancer both time-varying and in-homogeneous across the screening population[7]. To represent the risk of cervical cancer development, we consider three clinically actionable states, reflecting stages in disease initiation and progression. We label these states *normal*, *low-grade* and *high-grade*.

A normal state requires no additional exams before the next routine screening, while progression from normal to low-grade calls for closer follow-up – although low-grade lesions may spontaneously regress back to normal[8]. Progression from low-grade to high-grade requires immediate clinical action to prevent cancer. Each state is determined by the outcome of medical exams and corresponds to different risk-levels of disease development.

The NCCSP data is strongly skewed towards disease-free cases with more than 85 % of the individual results being normal and fewer than 5 % high-grade results. Due to the screening recommendations not being strictly adhered to in practice, the histories are irregular in time. This irregularity poses a significant challenge in prediction tasks if the time between the last examination and the time to predict amounts to several years (e.g. > 4 years). The panel in Fig. 1, illustrates these characteristics of the NCCSP data by showing to the left a Lexis diagram depicting screening histories, a histogram of screening intervals in the middle and a histogram of the proportion of female states in three age intervals to the right.

The Lexis diagram in Fig. 1 illustrates the scarcity in screening histories sampled from the NCCSP data, where the median number of exams is 6. The histogram of screening intervals shows that the time between exams varies from just over 1 month and up to almost 20 years, illustrating data irregularity. Finally, the proportion of states is changing with female age, containing about 0.87 normal results for females younger than 36 and up to 0.93 normals for 46–69+ year old females. This drift in the state distribution could be attributed to changes in female lifestyle and screening habits.

State of the art. Popular prediction models in biomedical applications include *logistic regression*[9] (LR), *random forest*[10] (RF) and *gradient tree boosting*[11] (GTB). Ensemble methods such as RF and GTB may be strong performers on imbalanced data[12] but neither of these models is typically used with time-dependent data. Popular models in time-series prediction tasks such as *long short-term memory*[13] (LSTM) networks expect regular and sufficiently sampled data. However, this is not the case with the NCCSP data, as described in the previous section.

An alternative to the LSTM, also capable of modelling cervical cancer data, is a continuous-time *hidden Markov model* (HMM) developed in a recent study[14] for the disease dynamics observed in cervical cancer screening data. The model was learned from a subset of NCCSP data and validated against a hold-out set by using the HMM as a stochastic simulator to derive Kaplan-Meier estimates. However, the study did not evaluate the HMM on risk prediction tasks or presented a method for generating such predictions from the model.

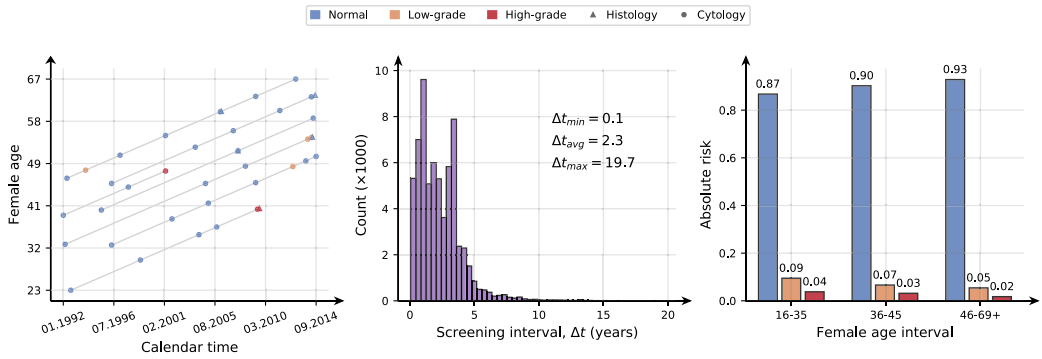

Figure 1. Cervical cancer screening data characteristics. Left: A Lexis diagram illustrating screening histories. Each history is depicted as a gray line spanning from the first to the last visit. Visits are indicated by a marker for the exam type (histology and cytology) and colored by the exam result. Middle: A histogram of the time between visits. Right: The proportion of female states (normal in blue, low-grade in orange and high-grade in red) in three age intervals.

A later work[15] introduced a *matrix factorization* (MF) framework using historical data for cervical cancer risk prediction along with a method for classifying the female state from the risk estimate. The MF has become a popular approach to dealing with scarce and irregular data. The proposed framework was compared to a *geometric deep learning* (GDL) model based on[16], and validated by the *probability of agreement*[17] between Kaplan-Meier estimates derived from model predictions and a hold-out set. Despite highlighting a heavy class imbalance in their data sample, the authors did not evaluate model calibration or state drift – factors that may explain and affect model performance.

In[18] the GDL approach was further adapted to handle the scarcity in the cervical cancer screening data. The method was evaluated in numerical experiments by predicting the future risk for individual females at a single randomly chosen time point. To extend methods for personalized risk prediction even further, incremental learning may be incorporated, allowing the models to update risk estimate after more data is available in the future.

Contribution. This is the first paper comparing several different machine learning methods for cervical cancer risk estimation, focusing on methods for incremental learning from longitudinal data, the impact of data imbalance on model estimates and classification with a time-varying state distribution. Specifically, we compare methods based on HMM, MF and GDL, as well as LR, RF and GTB. Our motivation for including the HMM, GDL and MF is that they were developed to handle scarce and irregular data for cervical cancer screening applications, and we further adapt these herein to incremental learning for longitudinal risk estimation. Moreover, to handle problems with both imbalance and temporal changes in the state distribution, we extend the classification method from[15] to separate classifiers over different female age intervals. To evaluate their ability to predict the next exam results over time, we compare absolute risk curves derived from model predictions and a hold-out set of screening data to assess model calibration against the trend in the time-varying risk.

The rest of this paper is organized as follows. In "Predicting the risk of cervical cancer development" we outline the risk estimators that are based on extensions to HMM, MF and GDL. "Numerical experiments" section describes the numerical experiments and discuss the results on NCCSP data, followed by a conclusion and outline of future work in "Conclusions and future work".

Predicting the risk of cervical cancer development

We represent a cervical cancer screening history with data recorded at times $t_0 < t_1 < \cdots < t_j$ as a set of tuples $\boldsymbol{y}_{t_j} = \{(t_i, \rho_{t_i}, x_{t_i})\}_{i=0}^{j}$. The history includes time points t_i representing the female age at visit i when she was measured with medical exam ρ_{t_i} to be in state $x_{t_i} = s$. The potential female states $s \in S$ are numerically encoded with $s = 1$ for normal, $s = 2$ for low-grade and $s = 3$ for high-grade.

To estimate the individual future risk of cervical cancer, we assume that we know her screening history up to some time t_j. The predicted risk at a later time point $\hat{t} > t_j$ is expressed as the triple of conditional probabilities $p(x_{\hat{t}} = s|\boldsymbol{y}_{t_j}), s = 1, 2, 3$.

In the following sections we provide a detailed description of how we extend existing frameworks based on MF, GDL and HMM to incremental learning for longitudinal predictions. For the LR, RF and GTB predictors we use the implementations from publicly available software[19].

Matrix factorization. The matrix factorization (MF) risk estimate as we define it herein is based on the *Shifted Weighted Convolutional Matrix Factorization* (SWCMF) from[15]. The SWCMF assumes that the discrete observed states are possibly inaccurate measurements of a continuous *latent state*, evolving slowly with time for each female. The MF risk estimator requires that we derive such latent state profiles from a hold-out set of screening histories before we can use it for predictions.

By organizing all the states x_{t_i} from a screening history according to female age t_i we obtain a scarce longitudinal vector $\mathbf{z}$ spanning $T \geq t_j$ years. Here T is the maximum female age in the data and the times $t_0 \leq t_i \leq T$ for when the female state was measured correspond to the observed entries in $\boldsymbol{z}$. Combining N such state vectors gives a partially observed matrix $\boldsymbol{Z}$.

With the SWCMF model from[15], we estimate a complete matrix $\widehat{\boldsymbol{M}} \in \mathbb{R}^{N\times T}$ of latent state profiles using the observed states in $\boldsymbol{Z}$ and the medical exam type data. Each row $\widehat{\boldsymbol{M}}_n$ in $\widehat{\boldsymbol{M}}$ corresponds to a continuous latent profile estimated from $\boldsymbol{Z}_n$, and these profiles are used to construct the MF risk estimator.

As in[15], we assume the probability of observing a female in state $x_{t_i} = s$ at time t_i given the latent state m_{t_i} is given by

$$p(x_{t_i} = s|m_{t_i}) = C \exp\left((m_{t_i} - s)^2/2\sigma^2\right) \tag{1}$$

where $C = C(m_{t_i})$ is a normalizing factor. Here we estimate σ using the same MLE procedure as in[15]. To estimate the risk of some female with screening history $\boldsymbol{y}_{t_j}$ being in state s at time $\hat{t} > t_j$, we consider the posterior predictive distribution

$$\begin{aligned} p(x_{\hat{t}} = s|\boldsymbol{y}_{t_j}) &\propto \int p(x_{\hat{t}} = s|\boldsymbol{m}, \boldsymbol{y}_{t_j}) p(\boldsymbol{m}|\boldsymbol{y}_{t_j})\, d\boldsymbol{m} \\ &= \int p(x_{\hat{t}} = s|\boldsymbol{m}) p(\boldsymbol{m}|\boldsymbol{y}_{t_j})\, d\boldsymbol{m}. \end{aligned} \tag{2}$$

In (2) we assume $y_{\hat{t}}$ is conditionally independent from $y_{t_0}, \ldots, y_{t_j}$ so the probability of observing state s when given history $\mathbf{y}$ and latent profile $\mathbf{m}$ becomes $p(x_{\hat{t}} = s|\boldsymbol{m})$. Using Bayes' rule $p(\boldsymbol{m}|\boldsymbol{y}_{t_j}) \propto p(\boldsymbol{y}_{t_j}|\boldsymbol{m})p(\boldsymbol{m})$ we get

$$p(x_{\hat{t}} = s|\boldsymbol{y}_{t_j}) \propto \int p(x_{\hat{t}} = s|\boldsymbol{m}) p(\boldsymbol{y}_{t_j}|\boldsymbol{m}) p(\boldsymbol{m})\, d\boldsymbol{m} \tag{3}$$

In (3) the latent risk prior $p(\boldsymbol{m})$ is unknown so $p(\boldsymbol{m}|\boldsymbol{y}_{t_j})$ is really intractable but following the variational approximation approach we may use $\widehat{\boldsymbol{M}}$ to approximate $p(\boldsymbol{m}|\boldsymbol{y}_{t_j})$. Thus, we can approximate (3) with

$$\widehat{p}(x_{\hat{t}} = s|\boldsymbol{y}_{t_j}) \propto \sum_{n=1}^{N} p(x_{\hat{t}} = s|\widehat{M}_{n,\hat{t}}) \widehat{p}(\boldsymbol{y}_{t_j}|\widehat{\boldsymbol{M}}_n). \tag{4}$$

In (4) we compute $p(x_{\hat{t}} = s|\widehat{M}_{n,\hat{t}})$ from (1) and the data likelihood as

$$\widehat{p}(\boldsymbol{y}_{t_j}|\widehat{\boldsymbol{M}}_n) = \prod_{i=0}^{j} C(\widehat{M}_{n,t_i}) \exp\left(\frac{(\widehat{M}_{n,t_i} - x_{t_i})^2}{2\sigma^2}\right),$$

Moreover, assuming data from a visit at time $t_{j+1} > t_j$ is added to history $\boldsymbol{y}_{t_j}$, we can recursively update the data likelihood by

$$\widehat{p}(\boldsymbol{y}_{t_{j+1}}|\widehat{\boldsymbol{M}}_n) \propto \widehat{p}(\widehat{\boldsymbol{M}}_n|\boldsymbol{y}_{t_j}) C(\widehat{M}_{n,t_{j+1}}) \exp\left(\frac{(\widehat{M}_{n,t_{j+1}} - x_{t_{j+1}})^2}{2\sigma^2}\right)$$

This recursive update was not described in[15] and allows us to do efficient adaptive learning by re-estimating the risk when more data is available.

Geometric deep learning. An alternative to using the SWCMF model[15] for estimating latent state profiles is to use a *geometric deep learning* (GDL) approach based on[16]. We define the GDL risk estimate based on latent profiles derived with GDL and using (4) for risk predictions. To estimate latent profiles, the GDL leverages two similarity graphs where one encode similarities between screening histories and the other represents the temporal dependency of results. When estimating the latent state profiles, GDL use these graphs to determine the structure of the profiles.

In[15] the authors used a k-nearest neighbour (NN) graph linking together similar histories in addition to a sequential graph for the temporal dependencies. In the k-NN graph, each node represents a screening history that is connected to k other most similar histories, where the similarity between histories is determined by some pre-defined measure. A potential drawback of the k-NN graph is that it each node has to have exactly k connections – even if one node is quite dissimilar from the others.

In this paper, we follow[18] in constructing a graph with a variable number of connections for each node. This graph is learned directly from the data under a smoothness constraint where we assume that certain screening histories exhibit strong similarities to each other. The resulting graph will then contain nodes connecting together histories that are alike in the results and time of visits. Moreover, we assume that the risk of cancer development does not change rapidly within a year and use this to construct the second graph for the temporal dependency of results. Using these two graphs with the GDL we obtain latent state profiles that are changing slowly in time and reflect the similarities between screening histories in the data. The GDL approach is similar to the MF estimate except that they use different constraints to characterize the latent state profiles.

Hidden Markov model. The HMM risk estimate that we compare to the MF and GDL estimates is based on an extension of the HMM described in[14] with a prediction module. In the HMM, each observed state is taken to originate from a discrete hidden state indicating the latent risk of cervical cancer development. Here we take the observed states to originate from some hidden states, labelled *normal*, *low-risk* and *high-risk* as in[14]. To define the HMM risk estimate, we first consider the probability

$$\alpha(h_{t_j}) = p(h_{t_j}|\mathbf{y}_{t_j})$$

of a female being in hidden state h_{t_j} at time t_j, conditioned on her screening history $\mathbf{y}_{t_j}$. We use this probability estimate to predict the risk at time $\hat{t} > t_j$. To compute $\alpha(h_{t_j})$ we initialize

$$\alpha(h_{t_0}) = p(h|t_0)p(x_{t_0}|h_{t_0}, \rho_{t_0})$$

Here, $p(h|t_0)$ is a prior over the hidden state at the time of the initial exam, and $p(x_{t_0}|h_{t_0}, \rho_{t_0})$ is the probability of the observed state conditioned on the medical exam and hidden state. The estimates for $p(h|t_0)$ and $p(x_{t_0}|h_{t_0}, \rho_{t_0})$ are available from the parameters of the HMM in[14]. To reach $\alpha(h_{t_i})$ for $t_i > t_0$, we use our previous estimate $\alpha(h_{t_{i-1}})$ to compute the recursion

$$\alpha(h_{t_i}) = p(x_{t_i}|h_{t_i}, \rho_{t_i}) \sum_{h_{t_{i-1}}} p(h_{t_i}|h_{t_{i-1}})\alpha(h_{t_{i-1}}). \tag{5}$$

The transition probabilities between hidden states $p(h_{t_i}|h_{t_{i-1}})$ are also given by the HMM parameters[14].

Having used (5) to obtain our estimate for the hidden state probabilities at time t_j, we predict the future risk at time $\hat{t} > t_j$ by approximating

$$p(x_{\hat{t}} = s|\mathbf{y}_{t_j}) \propto \int_{\rho_{\hat{t}}} \int_{h_{\hat{t}}} \int_{h_{t_j}} p(x_{\hat{t}} = s|h_{\hat{t}}, \rho_{\hat{t}})p(\rho_{\hat{t}}|h_{\hat{t}})p(h_{\hat{t}}|h_{t_j})\alpha(h_{t_j})\, dh_{t_j} dh_{\hat{t}} d\rho_{\hat{t}}. \tag{6}$$

The probabilities $p(\rho_{\hat{t}}|h_{\hat{t}})$ we derive from the Poisson intensity estimates presented in[14]. To incorporate more data and update the HMM risk estimate, like we do with MF and GDL, we first update the α estimate with (5) and then estimate the risk with (6).

Predicting the next state. Using any model to predict the risk of some female being in each state $s \in S$ gives a comprehensive overview of her risk. Predicting the exam result by classifying the female state from these risk estimates amounts to a multi-class classification problem. One approach to this task is to select the most probable state

$$\widehat{x}_{\hat{t}} = \arg\max_{s \in S} \widehat{p}(x_{\hat{t}} = s|\mathbf{y}_t).$$

However, this method often fails to predict the minority states because data imbalance shifts the risk inference and classification towards normal. We refer to this classification rule as the *default strategy* as it selects the most probable state without considering data imbalance.

An alternative to the default strategy is to consider state-specific probability thresholds $\{\delta_s \in (0, 1)\}_{s \in S}$ adapted to the skewed state distribution. To perform multi-class classification using these thresholds, we can construct a classification rule similar to[15] where for each state we evaluate

$$\widehat{p}(x_{\hat{t}} = s|\mathbf{y}_t) \geq \delta_s \quad \Longrightarrow \quad \widehat{x}_{\hat{t}} = s. \tag{7}$$

If condition (7) holds we predict $\widehat{x}_{\hat{t}} = s$. We first evaluate (7) for s being the high-grade state and then the low-grade state. If the condition is not satisfied for either of these states, we predict normal. This means that we prioritize predicting high-grade over low-grade, and low-grade over and normal as we in our application is more tolerant towards false positives than false negatives.

Furthermore, taking $\delta_s = \delta_s(t)$, we can adapt (7) to the label drift observed in our data by training a separate classifier for different female age intervals. Since the risk of HPV infection peaks in adolescence and early adulthood, and the risk of cervical cancer peaks in middle aged females[7], we choose three age intervals: 20–35, 36–45 and 46–69+ for our experiments. Moreover, choosing only three age intervals we aim to avoid overfitting as increasing the number of intervals would also increase the risk of overfitting the classifier in each interval. We refer to using (7) with time-dependent thresholds as the *adaptive strategy*.

To derive the probability thresholds $\delta(T_k)$ for each female age interval T_k, we maximize the K-category *Matthews correlation coefficient* (MCC)[20]. The MCC summarizes the confusion matrix in a single score

$$R_K = \frac{n_+ \times n - \sum_{s \in S} \hat{n}_s \times n_s}{\sqrt{(n^2 - \sum_{s \in S} \hat{n}_s^2) \times (n^2 - \sum_{s \in S} n_s^2)}}$$

to measure the quality of multi-class classifications. Here n is the total number of test samples, n_+ is the number of correct classifications, and n_s and $\hat{n}_s$ are the number of times where state s was the ground truth and was correctly predicted, respectively. Higher $R_K \in [-1, 1]$ means a more accurate classification. The thresholds $\delta(T_j)$ for age interval T_j are obtained by computing

Model	Normal	Low-grade	High-grade
MF	0.0830	**0.644**	**0.700**
HMM	0.0410	0.680	0.734
GDL	0.0430	0.683	0.863
GTB	**0.0220**	0.780	0.766
LR	0.0240	0.795	0.777
RF	0.0330	0.790	0.793

Table 1. Brier scores stratified by female states. The prediction models are matrix factorization (MF), hidden Markov model (HMM), geometric deep learning (GDL) gradient tree boosting (GTB), logistic regression (LR), and random forest (RF). Significant values are in bold.

$$\max_{\delta(T_k)\in(0,1)^S} R_K(\mathbf{x}, \widehat{\mathbf{x}}). \tag{8}$$

This maximization problem is solved by using the *differential evolution algorithm*[21].

Numerical experiments

The research for this study is approved by the South East Norway Regional Committee for Medical and Health Research Ethics (application ID: 11752). The health registry data used in this study does not originate from clinical trails and therefore the ethical committee granted this study with an exception from obtaining informed consent. All the research conducted herein accommodate the relevant guidelines and regulations.

In numerical experiments we study machine learning methods taking the individual screening history as input for cervical cancer risk prediction. The methods are based on: *hidden Markov model*[14] (HMM), *matrix factorization*[15] (MF), *geometric deep learning*[16] (GDL), *logistic regression*[9] (LR), *random forest*[10] (RF) and *gradient tree boosting*[11] (GTB). To predict the individual risk of cervical cancer development we use (6) for the HMM estimate, and (4) for MF and GDL. Although LR, GTB and RF treat each exam result as independent we facilitate adaptive learning by re-fitting the models with additional data, using the current estimate for model parameters as initialization.

As input to the HMM, MF and GDL predictors we provide all the data up to six months prior to the result we want to predict. For MF and GDL, the input data consist of female states and the corresponding time stamps, while the HMM also utilize exam type information. The input features to LR, RF and GTB combine the cumulative counts of each state conditioned on the exam type over time, together with the corresponding time stamps. For LR we used Z-scoring with parameters estimated from a hold-out set to normalize the features. To derive the latent state profiles used by the MF and GDL estimators, we leverage the exam results and the time stamps from the 10K histories sampled for our training set to construct the input matrix. Moreover, data from the training set is also used to fit LR, RF and GTB.

To simulate an environment for doing longitudinal adaptive learning with MF, HMM and GDL, we masked parts of the screening histories in the test set with a moving window. This way we mimic histories growing over time as a female has more exams. We start by revealing only the first 2 results to fit the estimators, and move forward in time to predict the 3rd result. For any history with more than 3 results, we repeatedly update the model by including the previous data point before we move to predict at the next result.

The impact of data imbalance on risk estimation. After fitting the models, we assess how well their risk estimates are calibrated using the Brier score. This score measures the agreement between the predicted risk $\widehat{p}$ and an indicator $o_{n,\hat{t}}(x_{n,\hat{t}} = s)$ for whether the result was actually $x_{n,\hat{t}} = s$. We compute the Brier score over N cases as

$$B = \frac{1}{N}\sum_{n=1}^{N}\left(\widehat{p}(x_{n,\hat{t}} = s|\mathbf{y}_n) - o_{n,\hat{t}}(x_{n,\hat{t}} = s)\right)^2.$$

In Table 1, we present Brier scores to evaluate the impact of class imbalance on model estimates. The scores were derived from model predictions aggregated over time and stratified by each ground truth state.

From Table 1 we see that we have lower Brier scores for normal states and higher scores for low-grade and high-grade states, which indicates that the prediction models are strongly biased towards the normal state. Thus, the model estimates are clearly affected by the skew in the state distribution. The GDL is especially poor at high-grade predictions but improves on low-grade, while MF is the best calibrated on high-grade followed by HMM.

Probability thresholding for risk classification. One way to alleviate biased probability estimates in classification tasks is to use a classification rule adapted to the data imbalance when converting probabilities into class labels. Using the adaptive thresholding technique from "Predicting the risk of cervical cancer development", we may also relax the effect of temporal drift in the state distribution by having a different classifier over female age intervals. In Fig. 2 we give the multi-class classification performance as R_K scores achieved with the adaptive and the default classification strategies.

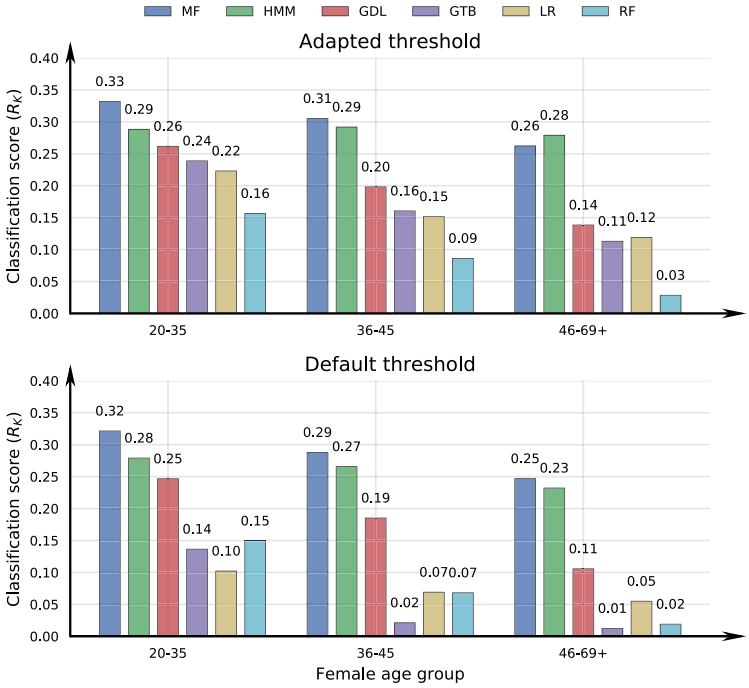

Figure 2. Classification performance as Matthews correlation coefficient (R_K) over female age intervals. The prediction models are matrix factorization (MF), hidden Markov model (HMM), geometric deep learning (GDL) gradient tree boosting (GTB), logistic regression (LR), and random forest (RF), combined with either the adapted or default probability threshold method from "Predicting the risk of cervical cancer development".

Comparing the R_K scores in Fig. 2 indicates that classification thresholds adjusted to class imbalance improves model performance and is the favourable method over the default strategy, especially with older females. The MF and HMM attains the strongest prediction performance. which is consistent with the model calibration estimates in Table 1, while the GDL, GTB, LR and RF performances decreases more over the age intervals.

Evaluating classifier performance. To assess how well the classifiers reflect the trends in the observed data, we compare absolute risk curves from hold-out data and longitudinal model predictions, using the strategy (either default or adaptive probability thresholds) improving on the classification scores in Fig. 2. Here we give absolute risk as the proportion of each state measured over some small time interval of about 10 months. In Fig. 3, we plot risk curves derived from test data and from model predictions. Each row in the panel figure corresponds to a different prediction model and there is one column for each state to more easily distinguish between the curves visually. Stippled vertical lines indicate the age intervals 20–35, 36–45 and 46–69+. Note that the scale on the y-axis differs between the normal/low-grade and high-grade plots to better illustrate the model fit. The colored regions illustrate the difference between the observed ($r(t)$) and predicted absolute risk ($\hat{r}(t)$) at time t. To quantify the relative deviation between the absolute risk curves, we define a performance indicator

$$\eta = \frac{\int |r(t) - \hat{r}(t)|\, dt}{\int r(t)\, dt}. \tag{9}$$

Ideally, $\eta = 0$, implying perfect classification, while miss-classifications cause the predicted curve to deviate from the test curve, giving $\eta > 0$.

The results in Fig. 3 indicate that the MF model is overall the most accurately calibrated against the trend in the reference curve from the hold-out registry screening data. The predictions from HMM and GDL improve over time, which may be attributed to an increasing amount of training data as older females have typically had more exams. The GTB, RF and LR estimates closely follow the reference curve for normal and low-grade but shows a large deviation in younger females which improves with older females.

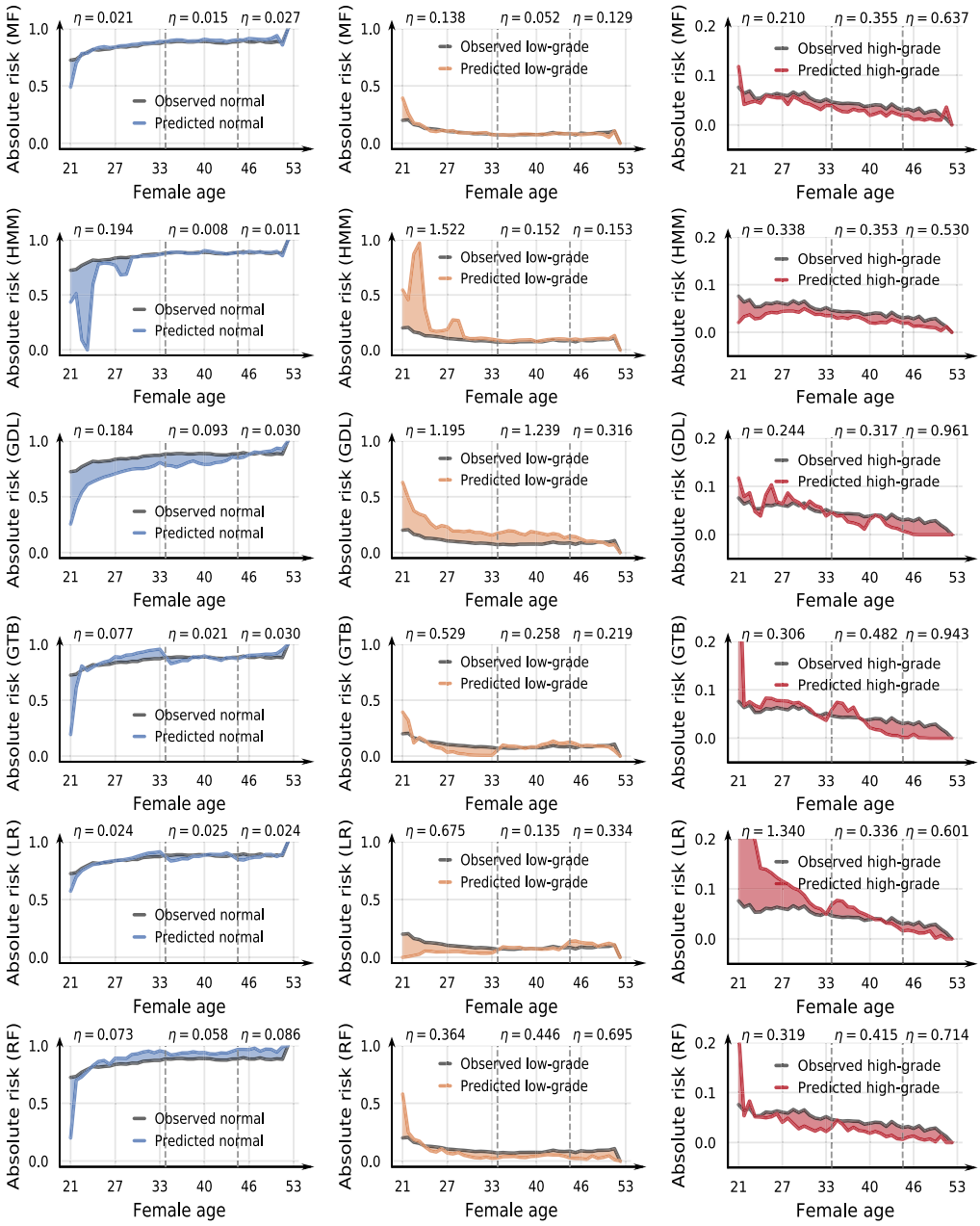

Figure 3. Absolute risk estimated from observed data and model predictions. The η score computed with (9) indicates model performance over female age intervals. The prediction models are matrix factorization (MF), hidden Markov model (HMM), geometric deep learning (GDL) gradient tree boosting (GTB), logistic regression (LR), and random forest (RF).

Conclusions and future work

Machine learning methods for more targeted risk stratification can have a high utility to existing cervical cancer screening programs shifting to more personalized screening recommendations. However, deriving such methods from cancer registry data is challenging due to strong class imbalance and a non-stationary data distribution. In this paper, we compare machine learning models based on matrix factorization (MF), hidden Markov model (HMM), geometric deep learning (GDL), logistic regression (LR), random forest (RF) and gradient tree boosting (GTB) in cervical cancer risk estimation, using population-level data from the *Cancer Registry of Norway*.

To define the risk estimators based on HMM, MF and GDL, we extend existing methods with incremental learning mechanisms for longitudinal risk prediction. Results from numerical experiments showed that all the models studied herein suffered from data skewness and were strongly biased towards disease-free results. To predict the individual risk of cancer development we trained separate classifiers adapted to data imbalance over separate female age intervals. Comparing absolute risk curves derived from model predictions and hold-out data showed promising results for matrix factorization to capture the time-varying trend in the observed risk from the data. This methods may thus be useful to improve cervical cancer risk stratification for more personalized screening. We are currently working to elucidate the ability of predictions models to correctly predict individual females using a different representation of model performance.

The methods used in this paper may also be applied to data from other types of mass-screening programs such as breast, colorectal and prostate cancer. In this paper, we focus on using only the routinely collected cervical cancer registry data as we see this to currently have more societal impact and utility for improving healthcare delivery. Expanding the models to include data from more recent screening technology with additional biomarkers and, eventually, individual HPV vaccination status has the potential to improve model performance. In future work we will combine female lifestyle information with registry screening data, believing that including more detailed information about each individual can improve the risk prediction accuracy.

Data availability

Due to individual privacy and ethical restrictions, the data used in this study are not publicly available. However, the data can be made available from the Cancer Registry of Norway pursuant the legal requirements mandated by the European GDPR.

Received: 31 January 2022; Accepted: 8 July 2022
Published online: 15 July 2022

References

1. Vaccarella, S. *et al.* 50 years of screening in the Nordic countries: Quantifying the effects on cervical cancer incidence. *Br. J. Cancer* **111**, 965–969 (2014).
2. Pedersen, K. *et al.* Advancing the evaluation of cervical cancer screening: Development and application of a longitudinal adherence metric. *Eur. J. Public Health* **27**, 1089–1094 (2017).
3. Perkins, R. B. *et al.* 2019 asccp risk-based management consensus guidelines for abnormal cervical cancer screening tests and cancer precursors. *J. Lower Genital Tract Dis.* **24**, 102 (2020).
4. Rothberg, M. B. *et al.* A risk prediction model to allow personalized screening for cervical cancer. *Cancer Causes Control* **29**, 297–304 (2018).
5. van der Waal, D. *et al.* Risk prediction of cervical abnormalities: The value of sociodemographic and lifestyle factors in addition to HPV status. *Prev. Med.* **130**, 105927 (2020).
6. Cohen, P. A., Jhingran, A., Oaknin, A. & Denny, L. Cervical cancer. *Lancet* **393**, 169–182. https://doi.org/10.1016/S0140-6736(18)32470-X (2019).
7. Schiffman, M. & Wentzensen, N. Human papillomavirus infection and the multistage carcinogenesis of cervical cancer. *Cancer Epidemiol. Prev. Biomark.* **22**, 553–560 (2013).
8. Castle, P. E., Schiffman, M., Wheeler, C. M. & Solomon, D. Evidence for frequent regression of cervical intraepithelial neoplasia-grade 2. *Obstet. Gynecol.* **113**, 18 (2009).
9. Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Open, 2017).
10. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
11. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
12. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **5**, 221–232 (2016).
13. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
14. Soper, B. C., Nygård, M., Abdulla, G., Meng, R. & Nygård, J. F. A hidden Markov model for population-level cervical cancer screening data. *Stat. Med.* **39**, 3569–3590 (2020).
15. Langberg, G. S. R. E. *et al.* Matrix factorization for the reconstruction of cervical cancer screening histories and prediction of future screening results (2021). Accepted for minor revision.
16. Monti, F., Bronstein, M. M. & Bresson, X. Geometric matrix completion with recurrent multi-graph neural networks. *arXiv preprint* arXiv:1704.06803 (2017).
17. Stevens, N. T. & Lu, L. Comparing Kaplan-Meier curves with the probability of agreement. *Stat. Med.* **39**, 4621–4635 (2020).
18. Gogineni, V. C. *et al.* Data-driven personalized cervical cancer risk prediction: A graph-perspective. In *2021 IEEE Statistical Signal Processing Workshop (SSP)* 46–50 (IEEE, 2021).
19. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
20. Gorodkin, J. Comparing two k-category assignments by a k-category correlation coefficient. *Comput. Biol. Chem.* **28**, 367–374 (2004).
21. Storn, R. & Price, K. Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* **11**, 341–359 (1997).

Author contributions

J.F.N. and G.S.R.E.L. conceived the experiments, G.S.R.E.L. conducted the experiments and M.G., V.N. and G.S.R.E.L. analyzed the results. All authors reviewed the manuscript.

Competing interests:
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to G.S.R.E.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Paper III

A weighted margin loss for treating imbalanced, overlapping and noisy data in cervical cancer risk prediction

Geir Severin R E Langberg, Markus Grasmair, Valeriya Naumova, Mari Nygård, Jan F Nygård

Submitted to *International Journal of Medical Informatics*, 2023.

III