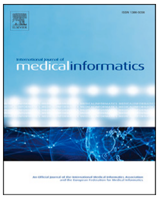
Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Cross-population evaluation of cervical cancer risk prediction algorithms

Severin Elvatun [a,*], Daan Knoors [b], Mari Nygård [c], Anneli Uusküla [d], Andres Võrk [e], Jan F. Nygård [f]

[a] Department of Registry Informatics, Cancer Registry of Norway, Ullernchausseen 64, 0379 Oslo, Norway
[b] Department of Registry Informatics, Cancer Registry of Norway, Norway
[c] Department of Research, Cancer Registry of Norway, Norway
[d] Department of Family Medicine and Public Health, University of Tartu, Estonia
[e] Institute of Economics, University of Tartu, Estonia
[f] Department of Registry Informatics, Cancer Registry of Norway, Department of Physics and Technology, UiT The Arctic University of Norway, Norway

ARTICLE INFO

Keywords:
Cervical cancer prevention
External validation
Risk prediction
Hidden Markov model
Matrix factorization
Neural network

ABSTRACT

Background: Cervical cancer is a preventable disease, despite being one of the most common types of female cancers worldwide. Integrating existing programs for cervical cancer screening with personalized risk prediction algorithms can improve population-level cancer prevention by enabling more targeted screening and contrive preventive healthcare innovations. While algorithms developed for cervical cancer risk prediction have shown promising performance in internal validation on more homogeneous populations, their ability to generalize to external populations remains to be assessed.

Methods: To address this gap, we perform a cross-population comparative study of personalized prediction algorithms for more personalized cervical cancer screening. Using data from the Norwegian and Estonian populations, the algorithms are validated on internal and external datasets to study their potential biases and limitations when applied to different populations. We evaluate the algorithms in predicting progression from low-grade precancerous cervical lesions, simulating a clinically relevant application of more personalized risk stratification.

Results: As expected, our numerical experiments show that algorithm performance varies depending on the population. However, some algorithms show strong generalization capacity across different data sources. Using Kaplan-Meier estimates, we demonstrate the strengths and limitations of the algorithms in detecting cancer progression over time by comparing to the trends observed from data. We assess their overall discrimination performance in personalized risk predictions by analyzing the accuracy and confidence in individual risk estimates.

Discussion and Conclusion: This study examines the effectiveness of personalized prediction algorithms across different populations. Our results demonstrate the potential for generalizing risk prediction algorithms to external populations. These findings highlight the importance of considering population diversity when developing risk prediction algorithms.

1. Introduction

Cervical cancer gradually develops from precursor lesions in a process that is usually initiated with a persistent infection with human Papillomavirus (HPV), [13]. The time to cancer development may take up to several years, and offers an opportunity to detect and treat the disease before becoming invasive. To promote early detection and to facilitate treatment, national programs for population-level cervical cancer prevention have been established in European countries, [1,10,11]. These programs follow national guidelines in recommending routine examination at regular intervals for early detection of cervical cancer development. Although successful in reducing cancer mortality, the current guidelines and recommendations for cervical cancer screening does not fully capture the heterogeneity of the individual risk. That is, the full target population is currently screened at regular time intervals to protect the sub-population of high risk individuals. As a consequence,

[*] Corresponding author.
E-mail address: sela@kreftregisteret.no (S. Elvatun).

https://doi.org/10.1016/j.ijmedinf.2023.105297
Received 29 September 2023; Received in revised form 11 October 2023; Accepted 12 November 2023
Available online 24 November 2023

regular screening of the entire target population will, for instance, lead to a high number of excessive exams and various financial costs, [2,12].

Turning to more personalized prevention strategies, the screening guidelines and recommendations are adapted to the individual need. Potentially, middle-aged women with a series of only normal results could be recommended less frequent screening, as they are at considerable lower risk than younger women with a history of several abnormal results. However, a challenge here is to identify the women who would benefit from closer follow-up and more intensive screening as well as those that can be recommended less frequent screening without compromising their protection. For instance, women detected with low grade cell changes can either progress to more severe cell changes, or the cell changes will regress back to a normal state, [15].

To provide women with low grade changes more informed recommendations, prediction algorithms for the individual risk of cervical cancer development can act as decision support. Previous studies, such as [6,7,9], have presented data-driven methods for individual risk prediction of cervical cancer development but have not specifically addressed their application to managing low grade results. Moreover, these studies have not conducted an external validation of their algorithms, assessing their ability to generalize across populations.

National screening programs for population-level cervical cancer prevention are established in Norway and Estonia. To administer the cervical cancer screening programs according to the national guidelines, the health authorities routinely collect information about the exam types (cytology, histology and HPV) and the corresponding exam results (normal, low grade, high grade and cancer) at each visit. Exam results are determined according to standardized guidelines by trained professionals via microscopic analysis of screening tests (cytology and HPV) or biopsy (histology). In Appendix A, Table 1, we compare distributional characteristics of the population-level datasets from Estonian and Norwegian cervical cancer screening programs. Although differing in size, the datasets have similar statistics. Routine screening in both populations produces a strong imbalance towards normal exam results and the number of exams per individual is typically scarce. The availability of the population-level dataset creates an opportunity to study data-driven approaches to more personalized cervical cancer prevention. Moreover, the similarities between the datasets suggest a potential for prediction algorithms to generalise across different populations.

To assess the ability of prediction algorithms to generalize across populations, we conduct a cross-population study of prediction algorithms for cervical cancer development. We evaluate the algorithms internally on data from the Norwegian cervical cancer screening population and use data from the Estonian population in the external validation. The algorithms are evaluated based on their ability to detect progression from low grade changes in a comparative study. We propose adjustments of the algorithms in [6,7,9] to improve their time-varying risk estimates, and compare their confidence and correctness in risk predictions, and their ability to detect more severe cell changes over time.

2. Methods

To predict the individual risk of a woman developing cervical cancer, we utilise historical data from her previous exams. The information that is available from each exam is the routinely collected data in population-level cervical cancer screening, namely the age, the exam type (cytology, histology and HPV) and the corresponding exam result (normal, low grade, high grade and cancer).

In our application, we are interested in predicting the risk of disease progression after having detected a low grade result. This means that we consider only women with at least one low grade result in our analyses. Moreover, we assume that each woman has previously had at least three exams (including one low grade) to provide the prediction algorithms with a minimum amount of input information.

2.1. Prediction algorithms

We consider in total four prediction algorithms. Comparing multiple algorithms also allows for evaluating whether the results from cross-population comparisons are algorithm-specific. Two of these algorithms are based on variants of a hidden Markov model (HMM) based on [14] and [9]. The third algorithm is an extension of a matrix factorization (MF) approach [7]. The fourth algorithm is based on recurrent neural networks (RNN), which are commonly applied in sequence modeling but may be unsuited for the irregularly sampled screening data.

The prediction algorithms provide estimates on the individual risk of cervical cancer development based on the information in the exam history of a woman. We represent a full exam history with the last exam at age t as $\mathbf{y}_t = \left\{ (t_i, \rho_{t_i}, x_{t_i}) \right\}_{t_i = t_0}^{t}$. Here, ρ is the exam type and x_t is the corresponding exam result. Note that the exam ages $t_0 \leq t_i \leq t$ usually vary considerably between the women. The two HMMs and MF algorithms are designed for more general applications than low grade management, expressing their risk estimates in terms of the posterior probabilities of exams result at $\hat{t} > t$. However, herein, we study only prediction estimates in scenarios where x_t is a low grade result and we want to predict the result at $\hat{t}$.

2.1.1. Hidden Markov model

We define the *hidden Markov model* (HMM) risk estimate as an extension of the method presented in [6], based on [14]. The HMM estimate approximates the posterior marginals for the next exam result based on an assumed underlying hidden state, indicating the latent risk of cervical cancer development. Compared to [6], we extend the method used to estimate the parameters for the risk estimator to account for more of the temporal variations in the parameters.

The HMM prediction estimate for having exam result x at age $\hat{t} > t$ conditioned on the individual exam history up to t is given by

$$p(x_{\hat{t}} = x \mid \mathbf{y}_t) \propto \sum_{\rho_{\hat{t}}} \sum_{h_{\hat{t}}} p(x_{\hat{t}} = x \mid h_{\hat{t}}, \rho_{\hat{t}}) p(\rho_{\hat{t}} \mid h_{\hat{t}}) p(h_{\hat{t}} \mid \mathbf{y}_t). \tag{1}$$

We marginalize over the exam type and hidden state since this information is unknown in advance of the prediction. The probabilities $p(x_{\hat{t}} = x \mid h_{\hat{t}}, \rho_{\hat{t}})$ and $p(\rho_{\hat{t}} \mid h_{\hat{t}})$ are obtained from [14]. We estimate the next hidden state conditioned on the exam history, $p(h_{\hat{t}} \mid \mathbf{y}_t)$, by marginalizing over the hidden states

$$p(h_{\hat{t}} \mid \mathbf{y}_t) = \sum_{h_t} p(h_{\hat{t}} \mid h_t) p(h_t \mid \mathbf{y}_t). \tag{2}$$

Here, $p(h_t \mid \mathbf{y}_t)$ is the conditional predictive posterior distribution of being in hidden state h_t at time t. This estimate is initialized at time t_0 as $p(h \mid t_0) p(x_{t_0} \mid h_{t_0}, \rho_{t_0})$, and recursively updated with

$$p(h_{t_i} \mid \mathbf{y}_{t_i}) = p(x_{t_i} \mid h_{t_i}, \rho_{t_i}) \sum_{h_{t_{i-1}}} p(h_{t_i} \mid h_{t_{i-1}}) p(h_{t_{i-1}} \mid \mathbf{y}_{t_{i-1}}). \tag{3}$$

For (2) and (3), we need to estimate the hidden state transition probabilities $p(h_{t_i} \mid h_{t_{i-1}})$ from the intensity parameters in [14]. The estimation method used in [6] does not account for temporal variations in the parameters. We therefore extend this approach herein by proposing adjustments to the prior estimates, described in Appendix B.

2.1.2. Hierarchical hidden Markov model

The *hierarchical hidden Markov model* (H-HMM) prediction algorithm builds on the model introduced in [9]. The H-HMM extends the HMM approach with the assumption that the screening population can be divided into two risk groups. These groups cover the women susceptible to progressing from low grade via a high grade to cancer, and the women that will only regresses back to normal.

The H-HMM prediction estimate is

$$p(x_{\hat{t}} = x \mid \mathbf{y}_t) \propto \sum_{\rho_{\hat{t}}} \sum_{h_{\hat{t}}} p(x_{\hat{t}} = x \mid h_{\hat{t}}, \rho_{\hat{t}}) p(\rho_{\hat{t}} \mid h_{\hat{t}}) q(h_{\hat{t}} \mid \mathbf{y}_t). \tag{4}$$

Different from [9], here we marginalize also over exam types, as this information is unknown at prediction time. The form of (4) resembles (1), except that for the dependency on the two risk groups $z \in \{z_0, z_1\}$, which yields

$$\begin{aligned} q(h_{\hat{t}} \mid \mathbf{y}_t) &= \sum_{h_t} \sum_{z} p(h_{\hat{t}} \mid \mathbf{y}_t, z) p(h_{\hat{t}} \mid h_t, z) p(z \mid \mathbf{y}_t) \\ &\propto \sum_{h_t} \sum_{z} p(h_{\hat{t}} \mid \mathbf{y}_t, z) p(h_{\hat{t}} \mid h_t, z) p(\mathbf{y}_t \mid z) \pi(z). \end{aligned}$$

We apply the procedure described in Appendix B also to estimate the conditional transition probabilities $p(h_{\hat{t}} \mid \mathbf{y}_t, z)$ from the intensity parameters in [9]. The predictive distribution of the model index $p(z \mid \mathbf{y}_t)$ is obtained from a prior estimate $\pi(z)$, which we treat as a hyperparameter. Specifically, we use $\pi(z_1) = 0.8$, in our numerical experiments in Section 3. We derive $p(\mathbf{y}_t \mid z)$ and $p(h_{\hat{t}} \mid \mathbf{y}_t, z)$ using the well-known forward-backward algorithm, [9].

2.1.3. Matrix factorization

An alternative to the HMMs is the *matrix factorization* (MF) prediction algorithm, which adapts the method presented in [6] to also utilise exam type information for risk estimation. Rather than assuming a fixed set of underlying states like the HMMs, the MF assumes the observed exam result is a potentially inaccurate measurement from a continuously evolving and time-varying latent risk profile for each woman. The probability of observing exam result x_t given the latent risk ψ_t is assumed to take the relationship

$$p(x_t \mid \psi_t) = c_t \exp(-\theta(x_t - \psi_t)^2).$$

Here, c_t is a normalizing factor, and $\theta > 0$ is a reliability parameter for the estimate, which we estimate from data, as described in Appendix C.

Here, we extend the probability of observing x_t from ψ_t to also depend on the exam type ρ_t. The conditional probability of x_t is thus decomposed as

$$p(x_t \mid \psi_t, \rho_t) = p(x_t \mid \psi_t) p(x_t \mid \rho_t).$$

The estimates for $p(x_t \mid \rho_t)$ are taken from [14]. Since the true latent risk profiles are unknown, we use a hold-out set of N exam histories to estimate a set of profiles $\mathbf{\Psi}$ that we use as proxies. We provide further details on how we estimate the risk profiles in numerical experiments in Section 3.2. Given $\mathbf{\Psi}$, the MF risk estimate is

$$p(x_{\hat{t}} = x \mid \mathbf{y}_t, \mathbf{\Psi}) \propto \sum_{\rho_{\hat{t}}} \sum_{n=1}^{N} p(x \mid \Psi_{n,\hat{t}}, \rho_{\hat{t}}) \prod_{t} p(x_t \mid \Psi_{n,t}, \rho_t). \tag{5}$$

Compared to [6], we introduce in (5) a marginalization over the exam types, adjusting the risk estimate to measurement uncertainties.

2.1.4. Recurrent neural network

As an alternative approach we implement a *recurrent neural network* (RNN) to predict the probabilities of regression and progression from a sequence of exam results [8]. That is, compared to the HMM, H-HMM and MF, the RNN outputs only the conditional probabilities of regression and progression. The input to the RNN consists of sequences with the age, the exam type and the corresponding exam result. In addition, we created an input feature from the time between exam results. Besides a traditional RNN, we also implemented its gated variants (i.e. LSTM and GRU) and select the one that performs best on each respective dataset. In the following sections we simply will refer to them as RNN.

While an RNN is frequently used to model regularly sampled and fixed length sequences, the data from cervical cancer screening contains variable length screening histories, with irregular time intervals between exam results. To handle the variable length histories, we used a packing approach, concatenating all histories and recording the indices of the start and end of each sequence. Further details on how we develop the RNN is specified in Section 3.2 of numerical experiments.

2.2. Performance evaluation

To assess the ability of an algorithm to predict cervical cancer development, we define an event as when a low grade result is followed by either another low grade, a high grade or a cancer result. The prediction accuracy of the algorithms is evaluated in terms of their ability to predict risk estimates and to classify events.

Risk estimation The risk estimate of having or not having an event at age $\hat{t}$ is derived from the conditional probability $p(x_{\hat{t}} = x \mid \mathbf{y}_t)$ of having a normal result x. From our data we know whether or not an event occurs at $\hat{t}$, and we use $\epsilon_{\hat{t}}$ to represent the ground truth outcome, i.e., event or no event. The probability estimate for the correct outcome $\epsilon_{\hat{t}}$ is thus

$$p(\epsilon_{\hat{t}} \mid \mathbf{y}_t) = \begin{cases} 1 - p(x_{\hat{t}} \mid \mathbf{y}_t) & \text{if event at } \hat{t} \\ p(x_{\hat{t}} \mid \mathbf{y}_t) & \text{if no event at } \hat{t}. \end{cases}$$

To quantify the confidence and correctness in a risk prediction, we consider the probability margin of correctly predicting the observed outcome, as defined by the δ score

$$\delta(\epsilon_{\hat{t}}) = p(\neg\epsilon_{\hat{t}} \mid \mathbf{y}_t) - p(\epsilon_{\hat{t}} \mid \mathbf{y}_t). \tag{6}$$

The δ score was first introduced in [3] with the interpretation that the model is confident about predicting the correct outcome if $\delta(\epsilon_{\hat{t}}) \approx -1$, since $p(\epsilon_{\hat{t}} \mid \mathbf{y}_t) \gg p(\neg\epsilon_{\hat{t}} \mid \mathbf{y}_t)$. Conversely, the model may be confident about predicting the incorrect outcome, in which case $p(\epsilon_{\hat{t}} \mid \mathbf{y}_t) \ll p(\neg\epsilon_{\hat{t}} \mid \mathbf{y}_t)$ and $\delta(\epsilon_{\hat{t}}) \approx 1$. Finally, $\delta(\epsilon_{\hat{t}}) \approx 0$ if the model is unsure about the outcome, because $p(\epsilon_{\hat{t}} \mid \mathbf{y}_t) \approx p(\neg\epsilon_{\hat{t}} \mid \mathbf{y}_t)$. This interpretation of the δ score provides insights into both the confidence and correctness of risk predictions.

To conclude whether an individual risk prediction is correct, we compare it to a predefined risk acceptance threshold $-1 \leq \tau \leq 1$. That is, if $\delta(\epsilon) \leq \tau$, we classify prediction $p(\epsilon \mid \mathbf{y})$ as correct. The total number of correct predictions for a given τ yields the sample coverage

$$\phi_\tau \propto |\delta(\epsilon) : \delta(\epsilon) \leq \tau|.$$

Plotting $\phi_\tau \in [0, 1]$ over varying τ creates a sample coverage curve, illustrating the distribution of algorithm confidence and correctness. A rapid incline in this curve for lower τ suggests confidence in predicting the correct outcomes, while increasing ϕ_τ for larger τ indicates less confident predictions. The area under the sample coverage curve (AUC), denoted Φ, yields an overall measure of algorithm performance in risk predictions. Choosing $\tau \in [-1, 1]$ gives $\Phi \in [0, 2]$, where higher Φ signify more accurate predictions.

Event classification The sample coverage provides only an aggregated estimate of algorithm accuracy. To assess the algorithm ability to predict events over time, we derive Kaplan-Meier (KM) curves from observed events, $S(t)$, and the corresponding predicted events, $\hat{S}(t)$. To create $\hat{S}(t)$, we select a single risk threshold $\tau^\star$ from an exhaustive search to minimize the difference between the observed and the candidate predicted curves

$$\tau^\star = \arg\min_{\tau} \int \left| S(t) - \hat{S}_\tau(t) \right| dt.$$

By comparing the KM curves visually, we can reveal time-varying trends in over-estimation and under-estimation of events.

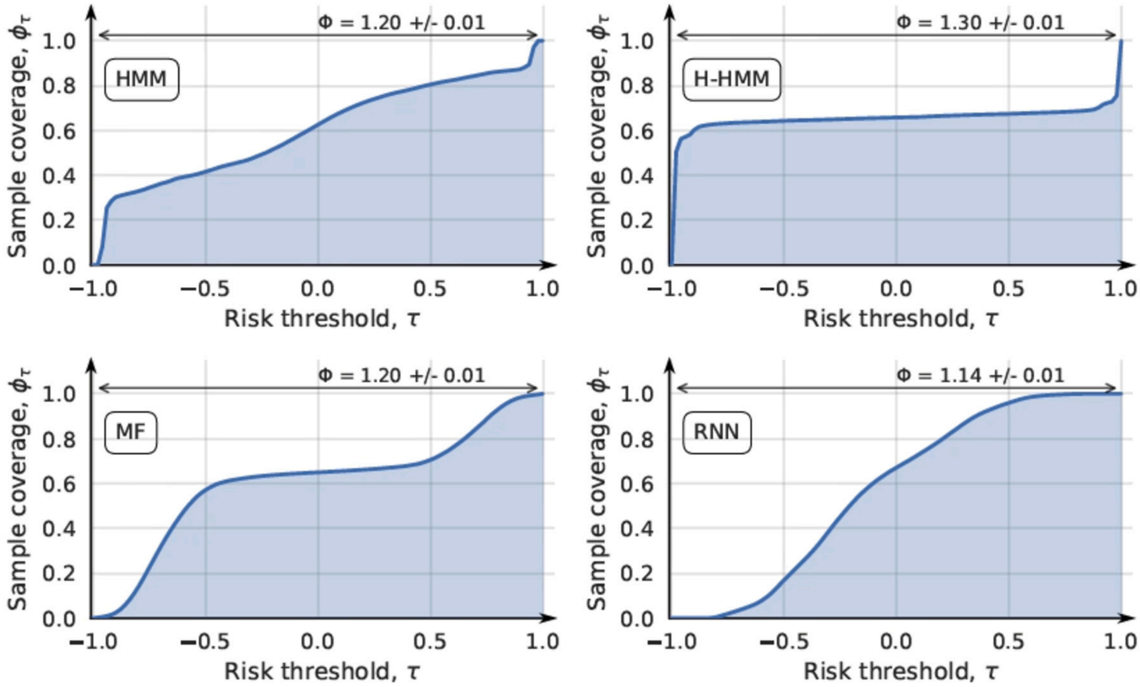


Fig. 1. Risk prediction over the Norwegian test set. Prediction algorithms are hidden Markov model (HMM), hierarchical HMM (H-HMM), matrix factorization (MF) and recurrent neural network (RNN). The area under the coverage curve is bounded to $\Phi \in [0, 2]$.

3. Numerical experiments

We compare four prediction algorithms based on a HMM, a H-HMM, MF and a RNN in a retrospective study predicting cervical cancer development from a low grade exam result. The algorithms are derived and validated internally on data from the Norwegian cervical cancer screening population, and validated externally on data from the Estonian screening population. To indicate the overall risk prediction accuracy, we use the AUC of the coverage curve, Φ, and we use Kaplan-Meier (KM) curves to illustrate prediction performance over time (see Section 2.2). Additional results on probability calibration and aggregate statistics on prediction performance are included in Appendix D and Appendix E (Tables 3 and 4), respectively. Confidence intervals around performance scores are obtained from ten bootstrap samples of the data.

3.1. Data

We use retrospective population-level datasets consisting of electronic screening records from the national programs for cervical cancer screening in Norway and Estonia. The Norwegian dataset spans from 1992 and 2020, while the Estonian results are from 2005 to 2020.

Focusing on a low grade management application, we exclude all women without at least one low grade results from our analyses. This reduce the number of individuals from 2,072,333 to 366,030 in the Norwegian dataset, and from 372,386 to 15,038 in Estonian dataset.

From both the Norwegian and Estonian datasets, we excluded any women younger than 16 years old[1] at the time of her first exam. Moreover, we excluded all women with less than four exam results in total, counting only results from cytology and histology exams. Summary statistics of the filtered datasets are provided in Appendix A, Table 2, and shows that the cohort we target in this study is slightly younger and have had more exams than the population average in Appendix A, Table 1. The fraction of low grades is larger in the filtered Estonian population, but both datasets contains similar proportions of events, as defined in Section 2.2.

3.2. Internal validation on the Norwegian data

We used the publicly available parameters from [14] and [9] to derive the HMM and H-HMM algorithms.

Remark 1. Both the HMM and the H-HMM are derived from data that may overlap with the test data we selected from the Norwegian population in this study. We had no way to determine which data had been previously used. Hence, our results from internal validation are subject to potential information leakage and should thus be interpreted with caution.

We randomly split the 366,030 Norwegian exam histories into 40% for training, 10% for validation and 50% for testing. Data splitting was performed before any pre-processing and model construction steps to avoid data leakage. Using the exam histories from the training set, we derive a set of latent risk profiles for the MF estimate, using the *shifted weighted convolutional* MF algorithm presented in [7]. The validation set was used to estimate thresholds for event classification, described in Section 2.2.

The training and validation sets were also used to optimize the RNN algorithm. It was trained with maximum 200 epochs with early stopping for convergence, using the ADAM optimizer and the binary cross entropy loss [5]. Hyperparameter optimization was performed using the Tree-structured Parzen Estimator (TPE) with over 200 trial candidates. The hyperparameter search space included one to five hidden layers, 2^0 to 2^6 hidden units, from 0–50% dropout, learning rates from 0.0001 to 0.1 and batch sizes from 2^4 to 2^9. The final RNN was a bi-directional LSTM with four hidden layers and two hidden units, 0.16% dropout, batch size 16 and learning rate of 0.0008.

In the following, performance statistics for risk and event prediction tasks are presented over the test set.

3.2.1. Risk prediction Norwegian data

We estimate the accuracy in event risk predictions by creating sample coverage curves, described in Section 2.2, and use the area under the curve, Φ, to compare algorithm performance. The coverage curves are plotted in Fig. 1.

[1] The age of consent is 16 years old in Norway.

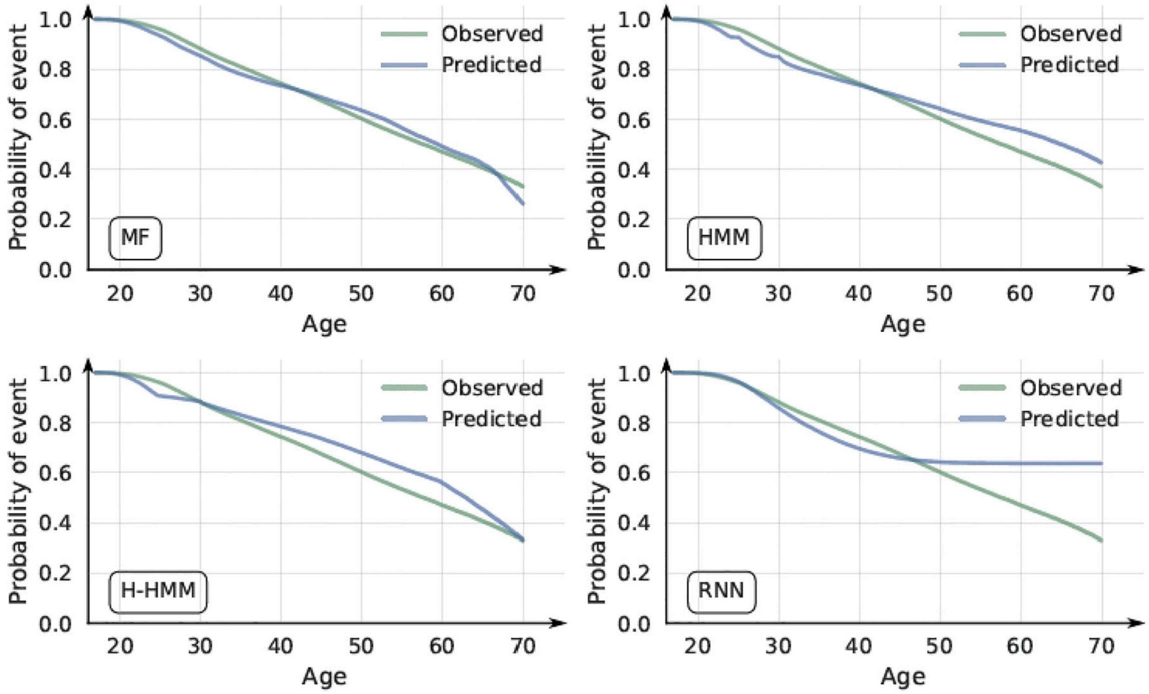


Fig. 2. Event classification over the Norwegian test set. Prediction algorithms are hidden Markov model (HMM), hierarchical HMM (H-HMM), matrix factorization (MF) and recurrent neural network (RNN). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

A rapid incline in the coverage curve for lower risk thresholds shows confidence in correct predictions, while a spike in the curve for higher thresholds suggest that the algorithms are confident in predicting the wrong outcome. Here, all prediction algorithms achieve similar performance. A distinguishing feature of the H-HMM is separating the screening population into disjoint risk groups, which may be beneficial to build confidence in the risk estimates as indicated by the slightly higher Φ score.

3.2.2. *Event classification Norwegian data*

To assess the ability of the algorithms to detect cervical cancer development over time, we compare in Fig. 2 the KM curves derived from the test data and the corresponding algorithm predictions. We classify risk predictions with the thresholds 0.38, 0.076, 0.20 and 0.40 selected for HMM, H-HMM, MF and RNN respectively. Smaller thresholds for H-HMM and MF can compensate for a tendency to underestimate the risk. Reliability curves illustrating algorithm calibration against the validation set are given in Appendix D. Error-based metrics for event classification performance on Norwegian test data is given in Table 3 in Appendix E.

Comparing the predicted and observed KM curves in Fig. 2 shows that the algorithms typically over-estimate the event rate for younger women and under-estimate the event rate for older women. The predictions for younger women may be affected by including histories with a minimum of only two exam results. Some of the variability in the HMM and H-HMM curves stem from time in-homogeneous model parameters, but the estimation method used herein reduced the artefacts.

3.3. *External validation on the Estonian data*

To assess their generalization capacity, we apply the prediction algorithms derived from the Norwegian data to data from the Estonian screening population for external validation. We randomly split the 15,038 Estonian exam histories into a validation set of 10% to estimate the risk classification thresholds, and we use the remaining 90% for performance evaluation.

3.3.1. *Risk prediction Estonian data*

We plot the sample coverage curves derived from risk predictions over the Estonian test set in Fig. 3, using the algorithms derived from the Norwegian data.

The Φ scores for external validation on the Estonian data strongly resembles the results from the internal validation on the Norwegian data (Fig. 1). However, the Estonian scores are in general slightly lower than for the Norwegian data, which is to be expected. While several outcomes can be predicted correctly in high confidence, an increased sample coverage for larger risk thresholds suggests that a distinct set of outcomes are harder to predict also in the Estonian data.

3.3.2. *Event classification Estonian data*

We also assess the ability of the algorithms to predict cervical cancer over time in the Estonian data in Fig. 4 and with to the Norwegian results in Fig. 2. Due to smaller sample size, the uncertainty profiles about the KM curves is larger here than for the Norwegian results.

To classify the Estonian risk predictions, we re-estimate the classification thresholds using the validation set. We selected risk thresholds at 0.39, 0.076, 0.20, 0.33 for the HMM, H-HMM, MF and RNN. Reliability curves illustrating algorithm calibration on the Estonian validation data are included in Appendix D. Error-based metrics for event classification performance on Norwegian test data is given in Table 4 in Appendix E.

The resemblance between the observed and predicted KM curves in Figs. 4 further suggest a capacity to generalize from the Norwegian data to predict the outcomes in the Estonian data. The trends observed for the Norwegian KM curves in Fig. 2 resembles the observed and predicted Estonian curves, where the event rate is slightly over-estimated for younger women and slightly under-estimated for older women.

4. Conclusion and future work

To reduce the high number of excessive exams in population-level cervical cancer screening, personalized risk prediction algorithms can be used to inform new guideline-based recommendations. Similarities in how these programs are implemented in European countries presents an opportunity to develop prediction algorithms generalizing to multi-

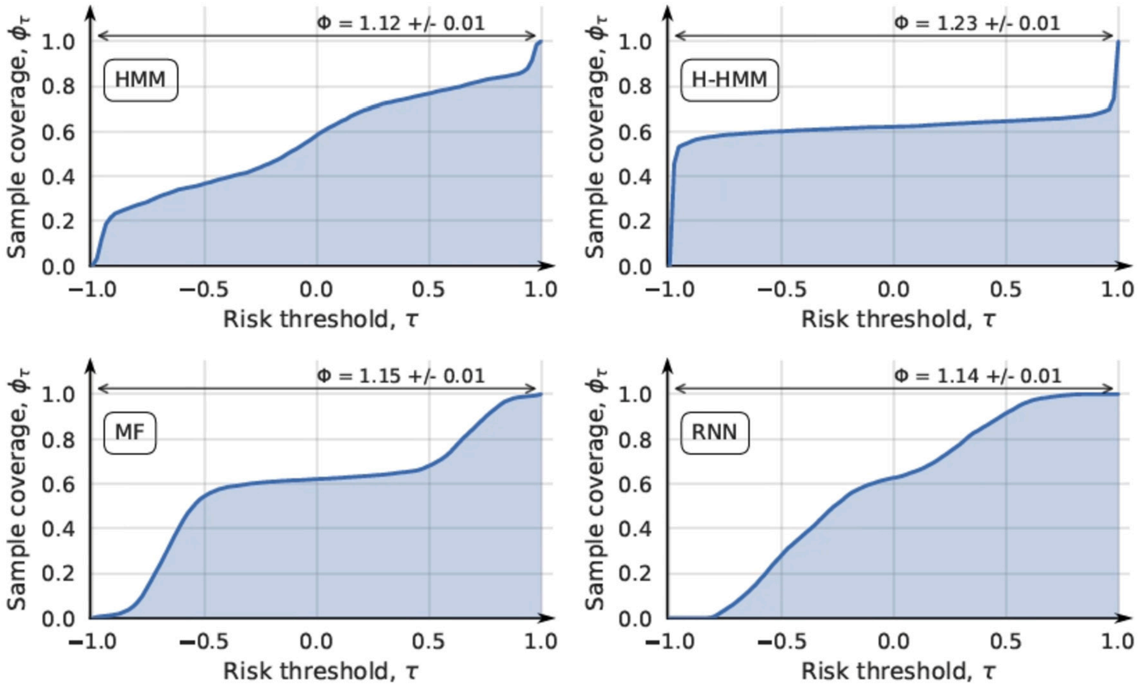


Fig. 3. Risk prediction over the Estonian test set. Prediction algorithms are hidden Markov model (HMM), hierarchical HMM (H-HMM), matrix factorization (MF) and recurrent neural network (RNN). The area under the coverage curve is bounded to $\Phi \in [0, 2]$.

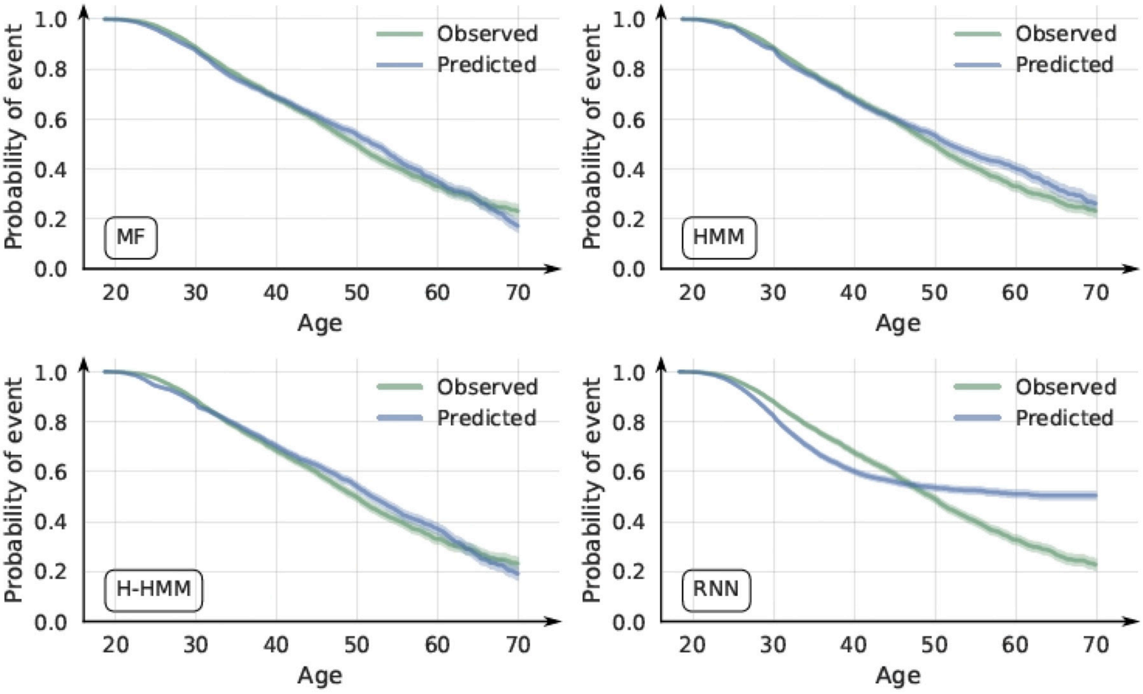


Fig. 4. Event classification over the Estonian test set. Prediction algorithms are hidden Markov model (HMM), hierarchical HMM (H-HMM), matrix factorization (MF) and recurrent neural network (RNN).

ple populations. To explore these opportunities, we conducted a cross-population comparative study of prediction algorithms, assessing their abilities to predict cervical cancer development and to generalize across populations, using data from the Norwegian and Estonian cervical cancer screening programs.

The prediction algorithms were derived from the Norwegian screening data, which was also used to validate the algorithms internally, and we used data from the Estonian screening population for external validation. Our numerical result indicate a strong ability in the algorithms to generalize from across populations. Although the algorithms did not achieve satisfactory discriminative performance, performance results from external validation are highly similar to internal validation results.

To further enhance the accuracy of the prediction algorithms, we recommend integration of additional data related to cervical cancer risk factors, such as HPV and smoking status. Including additional predictors may improve performance even for highly scarce screening data with few exam results per screening history.

Summary points

Background knowledge:

- Integrating existing programs for cervical cancer screening with personalized risk prediction algorithms can enable more targeted screening and contrive preventive healthcare innovations.
- Algorithms developed for population-level cervical cancer risk prediction have shown promising performance in internal validation on more homogeneous data, but their ability to generalize to external populations remains to be assessed.

Contributions:

- This study examines the effectiveness of personalized prediction algorithms across different populations, using data from the Norwegian and Estonian cervical cancer screening populations.
- Results from numerical experiments demonstrate the potential for generalizing risk prediction algorithms to external populations.

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) and Ethical review board in University of Tartu (protocol number: 3320/M-7, 21.12.2020) which waived the requirement to obtain informed consent. 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.

ChAMAI checklist

The authors confirm that this work is compliant with all the high-priority requirements of the ChAMAI checklist, [4].

CRediT authorship contribution statement

SE implemented and carried out the numerical experiments with the HMM, H-HMM and the MF algorithms. SE wrote the initial draft for the manuscript. DK implemented and ran the RNN experiments. All authors reviewed the manuscript.

Declaration of competing interest

The authors declare that they have no competing interests.

Data availability

The cervical cancer screening datasets used in this study can be made available from the cancer registries of Norway and Estonia pursuant the legal requirements mandated by the European GDPR, Article 6 and 9. The data and software code are not publicly available due to individual privacy and ethical restrictions.

Acknowledgements

The authors acknowledge the support for this project from the Research Council of Norway through the DeCipher project (IKTPLUSS grant 300034) and the EEA Grants (European Economic Area) project (EMP416).

Funding

This work is supported by the IKTPLUSS-program of the Research Council of Norway through the Decipher project (300034), as well as the grant EMP416 obtained from the EEA Grants (European Economic Area). The funding sources 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.

Appendix A

Table 1 and Table 2 provide summary statistics on the Norwegian and Estonian cohorts before and after applying the filtering criteria described in Section 2. Ethnic group information and socioeconomic status is not obtainable from the screening datasets in this study.

Appendix B

The HMM prediction algorithm presented in [6] is based on the time in-homogeneous HMM from [14]. The transition probabilities of this HMM are specific to disjoint age intervals. Specifically, the transition probability matrix $\mathbf{G}_k(t)$ at age t is derived from the transition intensity parameters $\mathbf{Q}_k$ specific to the age interval $k: \ \tau_k \leq t < \tau_{k+1}$. The approach to obtain the transition probabilities is based on the solution to the forward Komolgorov equations

$$\mathbf{G}_k(t) = \mathbf{G}(t_0)\exp(\mathbf{Q}_k \times t). \tag{7}$$

A common choice for the initial condition in (7), is $\mathbf{G}(t_0) = \mathbf{I}$ equal to the identity matrix. However, our preliminary experiments showed that $\mathbf{G}(t_0) = \mathbf{I}$ produced discontinuities in the prediction estimates for $k > 0$. One explanation for this observation is that the identity initial condition does not account for temporal dependencies between the age intervals. However, we found that adapting the initial condition to the specific age interval reduced variability in prediction estimates. Specifically, we use $\mathbf{G}_k(t_0) = \mathbf{I}$ for $k = 0$, and for $k > 0$ we defined

$$\mathbf{G}_k(t_0) = \prod_{j=1}^{i} \mathbf{G}(\tau_j) = \prod_{k=0}^{i-1} \exp(\mathbf{Q}_k \times (\tau_{k+1} - \tau_k)).$$

This adjusts the initial condition to a specific age interval, and the estimate for the adjusted transition probabilities becomes

$$\mathbf{G}_k(t) = \mathbf{G}_k(t_0)\exp(\mathbf{Q}_k \times (t - \tau_k)). \tag{8}$$

In this paper, we use (8) to estimate the transition probabilities for both the HMM and the H-HMM.

Appendix C

The error model for the MF algorithm is based on a normal distribution with mean time-varying parameters Ψ and a variance $\sigma^2 = 1/(2\theta)$. Let $\mathbf{Y} \in \mathbb{N}^{N \times T}$ be a matrix representation of N screening histories over T time points, as described in [7]. The likelihood of the screening data is given by

$$\begin{aligned} p(Y \mid \theta, \Psi) &= \prod_{(n,t)\in\Omega} \mathcal{N}(Y_{n,t}, 1/2\theta, \Psi_{n,t}) \\ &= \prod_{(n,t)\in\Omega} \frac{\sqrt{\theta}}{\sqrt{\pi}} \exp(-\theta(Y_{n,t} - \Psi_{n,t})^2) \\ &= \left(\frac{\theta}{\pi}\right)^{\frac{|\Omega|}{2}} \exp\left(-\theta \sum_{(n,t)\in\Omega} (Y_{n,t} - \Psi_{n,t})^2\right) \end{aligned}$$

Here, $|\Omega|$ denotes the total number of elements $(n,t) \in \Omega$. By taking the log of the likelihood we have

Table 1
Summary statistics of Estonian and Norwegian cohorts.

Statistic		Estonia	(%)	Norway	(%)
Number of women		372386	100	2072333	100
Women with low grade		15038	4.0	366030	17.7
Age at first exam	Min	16		16	
	Mean	44.5		36.5	
	Std	11.2		16.8	
	Max	101		103	
Exam counts	Min	1		1	
	Median	3		5	
	Max	37		50	
Exam results	Normal	1249176	95.0	12186340	93.1
	Low grade	22463	1.7	635496	4.9
	High grade	25689	1.9	264738	2
	Cancer	17058	1.4	7875	0.1
Exam types	Cytology	1232662	93.8	12795114	97.7
	Histology	81724	6.2	299335	2.3

Table 2
Summary statistics of the Estonian and Norwegian cohorts after applying the filtering criteria described in Section 3.1.

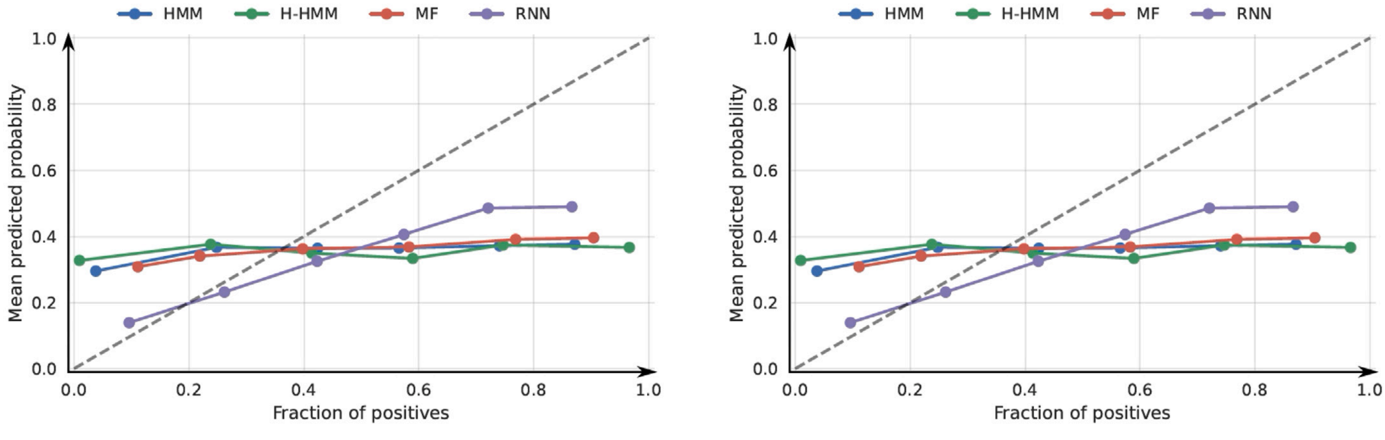
Statistic		Estonia filtered	(%)	Norway filtered	(%)
Number of women		11810		321419	
Number of events		4607	39.0	125093	38.9
Age at first exam	Min	16.4		16.0	
	Mean	32.8		29.5	
	Std	11.2		11.4	
	Max	87.7		89.9	
Exam counts	Min	4		4	
	Median	7		10	
	Max	28		50	
Exam results	Normal	61131	71.7	2726952	78.2
	Low grade	18178	21.3	585242	16.8
	High grade	5479	6.4	164325	4.7
	Cancer	428	0.6	2192	6.3
Exam types	Cytology	76390	89.6	3332798	95.8
	Histology	8826	10.4	145913	4.8

$$\begin{aligned}\ln p(Y \mid \theta, \Psi) &= \frac{|\Omega|}{2}(\log\theta - \log\pi) - \theta \sum_{(n,t)\in\Omega} (Y_{n,t} - \Psi_{n,t})^2 \\ &= l(\theta).\end{aligned}$$

By standard means of maximum likelihood estimation, we apply the derivative to $l(\theta)$ and solve for $dl/d\theta = 0$, which yields the MLE estimate

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

Appendix D



(a)

(b)

Fig. 5. Reliability curves illustrating the prediction algorithm calibration level against the 5a Norwegian and 5b Estonian validation sets.

Appendix E

Table 3
Error-based metrics for event classification performance on Norwegian test data. Confidence intervals were estimated from ten bootstrap samples over the test set.

Score	HMM	H-HMM	MF	RNN
Accuracy	0.587 ± 0.001	0.586 ± 0.001	0.576 ± 0.001	0.636 ± 0.001
Balanced accuracy	0.525 ± 0.002	0.514 ± 0.001	0.526 ± 0.002	0.569 ± 0.001
Sensitivity	0.346 ± 0.003	0.305 ± 0.002	0.381 ± 0.003	0.376 ± 0.002
Specificity	0.705 ± 0.001	0.723 ± 0.001	0.672 ± 0.001	0.762 ± 0.002
ROC AUC	0.546 ± 0.001	0.540 ± 0.001	0.534 ± 0.002	0.622 ± 0.001
Brier loss	0.546 ± 0.002	0.540 ± 0.001	0.534 ± 0.002	0.621 ± 0.001

⋆: Receiver operating characteristic

Table 4
Error-based metrics for event classification performance on Estonian test data. Confidence intervals were estimated from ten bootstrap samples over the test set.

Score	HMM	H-HMM	MF	RNN
Accuracy	0.537 ± 0.006	0.556 ± 0.004	0.545 ± 0.004	0.624 ± 0.006
Balanced accuracy	0.499 ± 0.006	0.520 ± 0.004	0.512 ± 0.004	0.597 ± 0.005
Sensitivity	0.358 ± 0.009	0.384 ± 0.009	0.391 ± 0.008	0.490 ± 0.009
Specificity	0.641 ± 0.005	0.656 ± 0.004	0.634 ± 0.004	0.703 ± 0.007
ROC* AUC	0.496 ± 0.007	0.513 ± 0.005	0.509 ± 0.006	0.651 ± 0.006
Brier loss	0.495 ± 0.007	0.513 ± 0.005	0.509 ± 0.006	0.651 ± 0.006

⋆: Receiver operating characteristic

References

[1] A. Anttila, G. Ronco, G. Clifford, F. Bray, M. Hakama, M. Arbyn, E. Weiderpass, Cervical cancer screening programmes and policies in 18 European countries, J. Med. Screen. 91 (5) (2004) 935–941, Nature Publishing Group.

[2] Renée M.F. Ebisch, Maroeska M. Rovers, Remko P. Bosgraaf, Helga W. van der Pluijm-Schouten, Willem J.G. Melchers, Petronella A.J. van den Akker, Leon F.A.G. Massuger, Ruud L.M. Bekkers, Evidence supporting see-and-treat management of cervical intraepithelial neoplasia: a systematic review and meta-analysis, BJOG Int. J. Obstet. Gynaecol. 123 (1) (2016) 59–66, Wiley Online Library.

[3] Severin Elvatun, Markus Grasmair, Valeriya Naumova, Mari Nygård, Jan F. Nygård, A weighted margin loss for treating imbalanced, overlapping and noisy data, Unpublished, 2023.

[4] Cabitza Federico, Andrea Campagner, The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies, Int. J. Med. Inform. 153 (2) (2014) 104510, Elsevier.

[5] Diederik P. Kingma, Ba Jimmy, Adam: a method for stochastic optimization, arXiv preprint, arXiv:1412.6980, 2014.

[6] G.S.R.E. Langberg, Jan F. Nygård, Vinay C. Gogineni, Mari Nygård, Markus Grasmair, Valeriya Naumova, Towards a data-driven system for personalized cervical cancer risk stratification, Nat. Portf. (2022).

[7] G.S.R.E. Langberg, Mikal Stapnes, Jan F. Nygård, Mari Nygård, Markus Grasmair, Valeriya Naumova, Matrix factorization for the reconstruction of cervical cancer screening histories and prediction of future screening results, BMC Bioinform. (2022).

[8] Larry R. Medsker, L.C. Jain, Recurrent neural networks, Des. Appl. 5 (2) (2001) 64–67, Nature Publishing Group.

[9] Rui Meng, Braden C. Soper, Herbert Kh Lee, Jan F. Nygård, Mari Nygård, Hierarchical continuous-time inhomogeneous hidden markov model for cancer screening with extensive followup data, Stat. Methods Med. Res. (2022) 9622802221122390, https://doi.org/10.1177/09622802221122390.

[10] J.F. Nygård, G.B. Skare, S.Ø. Thoresen, The cervical cancer screening programme in Norway, 1992–2000: changes in Pap smear coverage and incidence of cervical cancer, J. Med. Screen. 9 (2) (2002) 86–91, SAGE Publications Sage UK: London, England.

[11] Kristiina Ojamaa, Kaire Innos, Aleksei Baburin, Hele Everaus, Piret Veerus, Trends in cervical cancer incidence and survival in Estonia from 1995 to 2014, BioMed Central 18 (1) (2018).

[12] Kine Pedersen, Emily A. Burger, Suzanne Campbell, Mari Nygård, Eline Aas, Stefan Lönnberg, Advancing the evaluation of cervical cancer screening: development and application of a longitudinal adherence metric, Eur. J. Public Health 27 (6) (2017) 1089–1094, Oxford University Press.

[13] Mark Schiffman, Nicolas Wentzensen, Human papillomavirus infection and the multistage carcinogenesis of cervical cancer, Cancer Epidemiol. Prev. Biomark. 22 (4) (2013) 553–560.

[14] C. Braden Soper, Mari Nygård, Ghaleb Abdulla, Rui Meng, Jan F. Nygård, A hidden Markov model for population-level cervical cancer screening data, Stat. Med. 39 (2020).

[15] G.J. Van Oortmarssench, J.D.F. Habbema, Epidemiological evidence for age-dependent regression of pre-invasive cervical cancer, Br. J. Cancer 64 (3) (1991) 559–565, Nature Publishing Group.