NIHR | Applied Research Collaboration Kent, Surrey and Sussex

Using machine learning to predict 10-year risk of endometrial cancer: An External Validation using the CPRD Database

Rebecca Karkia
Senior Clinical Research Fellow and Doctoral student.
Royal Surrey NHS Foundation Trust
Brunel University.

NHS Royal Surrey NHS Foundation Trust

Brunel University of London

Public Summary:

Endometrial cancer is the fourth most common cancer of women in the United Kingdom. The largest risk factor for endometrial cancer is obesity but factor such as type two diabetes and polycystic ovarian syndrome are also risk factors. In 2022, 35% of the adult population were overweight and 29% were obese and as such it is difficult to accurately predict those at highest risk[1]. A risk prediction algorithm to identify those at highest risk of cancer development would allow safe and effective risk reducing measures to be offered for women.

In previously published work by Hutt et al, a machine learning algorithm which sought to predict ten-year risk of endometrial cancer[2]. Using 6 readily available pieces of medical data including body mass index and diabetes status, the algorithm was built and tested on a cohort of approximately 40,000 patients.

The algorithm on initial testing performed excellently suggesting that it would be a useful tool for risk prediction.
The next step in model evaluation was to use the same model to test a different cohort. This is known as external validation. If a model performs well at this step, then it is likely that the model is of use at a population level. The Clinical Practice Research Datalink (CPRD) which contains primary care records of approximately 10 % of the population of the United Kingdom was used. Over 3 million women were included in the cohort for external validation. Within this population, the incidence of endometrial cancer was low at 0.3%.

The algorithm's performance was evaluated using measures of external calibration. On final testing it appeared that the algorithm does not predict ten-year risk of endometrial cancer with a high enough degree of accuracy as to be implemented in clinical practice. Further model development is required.

Public and Community Involvement:

According to the James Lind Alliance, the top research priority as it pertains to cancers of the womb is:

"Is it possible to develop a personalised score which reflects a woman's individual risk of developing endometrial cancer?"

At present, there is no tool that is widely used in clinical practice to do this. For this reason, this was deemed to be an important research question and formed the basis of this study.

Given that this risk prediction algorithm has not been found to be effective enough for further use, there has been no additional public and community involvement.

It will be essential to understand the thought and feelings of potential users of a risk prediction tool before it is widely offered. Given that there are a number of interventions such as diet, exercise and certain medications which can lower risk of endometrial cancer. It is likely that a personalised risk score followed by individualised risk reduction measures would be widely accepted.

Dissemination:

At present, the findings from this study have not been published.

Impact and Implementation:

The vast majority of risk prediction algorithms which look promising on first testing, fail to perform adequately at external validation.

Neural networking as a tool for risk prediction in cancer has not been widely used. It is likely that this not the best machine learning tool for prediction of rare outcomes.

At present, conventional statistical measures appear to outperform machine learning in the landscape of endometrial cancer risk prediction. It is likely this will change in the future as machine learning and bioinformatic techniques are improving rapidly.

What next?

More work is needed to investigate how machine learning techniques and artificial intelligence can improved personalized risk prediction for endometrial cancer.

This research was supported by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration Kent, Surrey and Sussex. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.