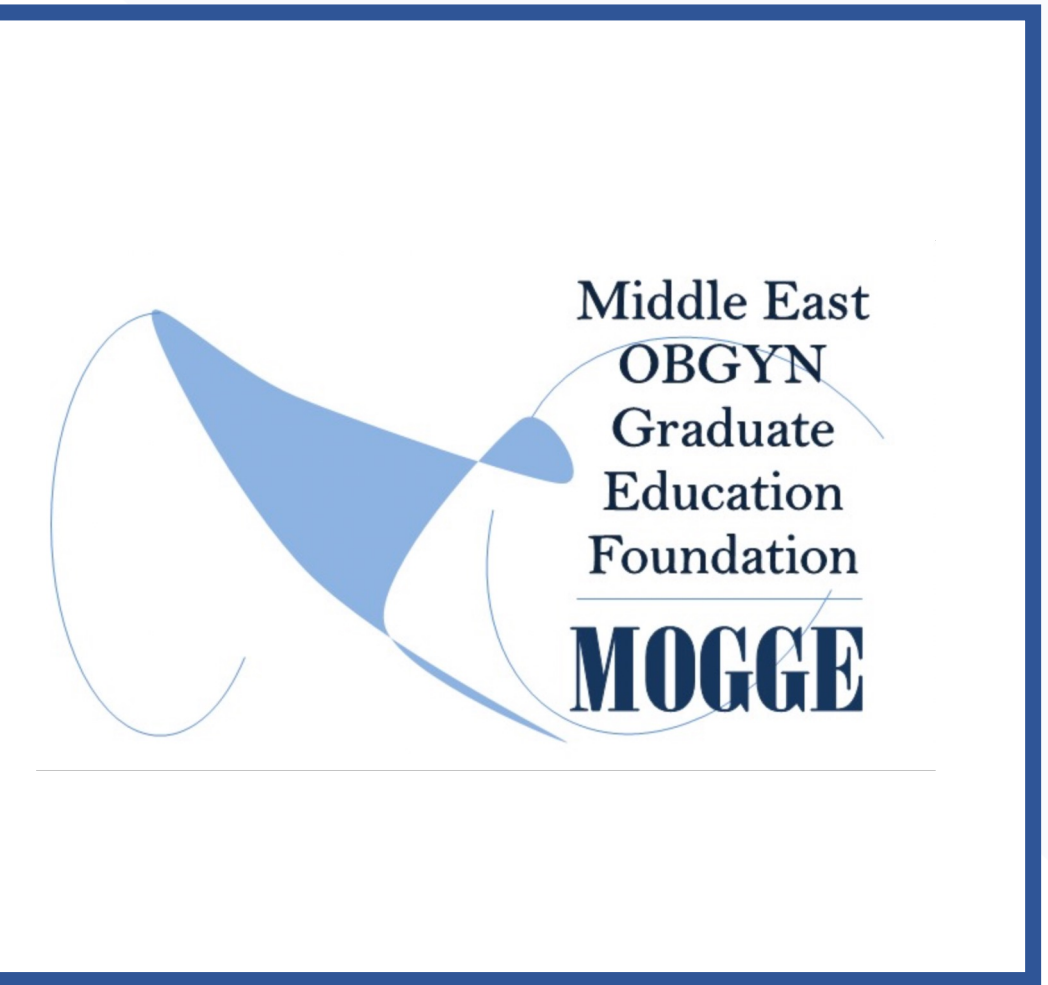




Endometrial cancer prediction model: A machine learning-based prognostic scoring system (Protocol)

Sherif Shazly, MBBCh, MSc¹

¹ Leeds Teaching Hospitals, Leeds, United Kingdom, Middle-East Obstetrics and gynecology graduate education (MOGGE) foundation



Abstract

Objective: To create a predictive model of endometrial cancer prognosis that considers patient and disease characteristics using machine learning approach, and to facilitate individualization of management by predicting the impact of presumed treatment plans on a specific patient survival. Thereby, the model could serve as an alternative to traditional cancer staging systems.

Design: This is a multicentre international retrospective study (ECID) including 9 European centres. Women who received treatment for endometrial cancer between January 2008 to December 2015, with a total follow-up of at least 3 years, will be considered in the study. Required variables will include patient age, parity, body mass index, ethnicity, medical comorbidities, preoperative disease characteristics (tumour size, extent of myometrial invasion, parametrial, cervical and vaginal involvement, iliac and paraaortic lymph node involvement, omental/peritoneal involvement, distant metastasis), staging, histopathology (histological type, grading, lymph vascular space invasion), family and personal history of cancer, treatment approach (surgical management including lymphadenectomy, brachytherapy, external pelvic radiation, chemotherapy, and hormonal therapy). Primary outcomes include overall survival. Disease free survival, recurrence rate, and postoperative complications comprise secondary outcomes. A machine learning model will be created using all patient and disease characteristics as independent variables (features). For model one, overall survival in months will be used as a dependent continuous variable (target) using machine learning regression algorithms. For model two, independent variables will include treatment variables in addition baseline characteristics, and the same target (overall survival) will be used. Models three and four will reproduce independent variables of model three and four to predict recurrence. Models will be tested using K cross validation. Machine learning models will be created using python 3.8®, through Anconda 3® interface (NCT04787159).

Results: Data will be described in means, medians, standard deviations, interquartile ranges, numbers and percentages. Model performance will be presented as area under curve, and prediction accuracy. A learning curve will be plotted to monitor machine learning progress. The magnitude of impact of each variable (feature) in the final score will be demonstrated as a graph.

Conclusion: The current study would provide a tool for individualized prediction of endometrial cancer prognosis and assessment of potential treatment plans per patient and disease characteristics.

Introduction

Endometrial cancer (EC) is the most common gynaecologic cancer in developing countries and is the sixth most diagnosed malignancy in women worldwide. In 2020, 417,367 new cases were diagnosed, and 97,370 EC-related mortalities were reported globally. Ideally, a cancer staging system would be designed to classify disease behaviour, indicate prognosis, and contribute to treatment decision. For EC, the International Federation of Gynaecology and Obstetrics (FIGO) staging has been adopted as the standard classification system since 1970. Whereas FIGO staging is reflective of disease prognosis, and stage III is associated with 5-year overall survival of 58%, compared to 91% with stage I disease, it is not inclusive of all prognostic factors, such as histologic type, pathological grade, and lymphovascular space invasion. In fact, disease grade, myometrial and cervical stromal invasion was associated with greater prognostic impact than pelvic lymph node metastasis. These factors would impact therapeutic decision-making and contribute to management plan over staging alone. This highlights the importance of individualised multi-factorial scoring system for assessment of disease prognosis and delineation of treatment. Machine learning is a subtype of artificial intelligence where machine progressively recognizes patterns that link variables to a specific outcome. In comparison to traditional statistics, machine learning establishes comprehensive individualised prediction models rather than association between individual variables and a clinical outcome.

Objectives

The objective of this study is to develop an EC scoring system, which predicts individualised prognosis based on patient demographics and disease characteristics using machine learning algorithms. In addition, the study aims at predicting prognosis in response to therapeutic plan, which would facilitate individualised treatment decision in women with EC.

Material and methods

Endometrial Cancer International Database (ECID) is a multicentric data collection project, which was launched for the purpose of the current study. ECID comprises the collaboration of 9 European cancer centres; women who were diagnosed with EC and received treatment in the contributing centres between January 1st, 2008, and December 31st, 2015, were considered eligible.

Inclusion criteria

- Women diagnosed with endometrial cancer between 2008 and 2015
- Adult women aged 18 or more years
- Women should be diagnosed and managed by the corresponding centre

Exclusion criteria

- Inadequate information and follow-up for at least 3 years
- Synchronous cancers
- Inadequate documentation of management/pathology of the disease

Data will be collected using an excel spreadsheet designed for this study or an online form (as per center preference). Data include patient age, parity, body mass index, ethnicity, smoking index, contraception method, medical comorbidities [coronary artery disease, diabetes on insulin, hypertension, chronic renal disease, chronic lung disease, thyroid dysfunction], preoperative imaging [tumor size, myometrial involvement (< 50% or > 50%), parametrial invasion, cervical involvement (radiological, clinical), vaginal involvement (upper third, middle third or lower third), iliac lymph nodes (number, largest diameter), paraaortic lymph nodes (number, largest diameter), omental/peritoneal involvement (isolated, disseminated), lung metastasis (single, multiple, largest nodule), pleural effusion, liver metastasis (single, multiple, largest nodule), ascites], staging, histopathology [histological type, grading, lymph vascular space invasion “LVSI”], genetics [positive family history, Lynch syndrome], treatment approach [surgical approach, hysterectomy BSO, omentectomy, lymph node management (sentinel lymph nodes, pelvic lymphadenectomy, para-aortic lymphadenectomy, lymph node sampling), brachytherapy, external pelvic radiation, adjuvant chemotherapy], and related outcomes (overall survival, disease free survival, and rate of distant metastases at 1 year, 3 years, and 5 years, pelvic organ functioning, postoperative complications, incidence of lymphedema, tumor-related deaths, frequency of residual disease on definitive surgical specimens). Data will not include any identifiable information. With the data gathered from this database, the investigators could further demonstrate different aspects of EC management and prognosis.

Model	Variables included
Model I	Preoperative model which will include patient baseline demographics (e.g., age, ethnicity, body mass index) and disease characteristics (e.g., pathological type, preoperative imaging details, tumour grade, LVSI)
Model II	Probability calculated from model I + management details (surgical management, radiotherapy, chemotherapy, hormonal treatment)
Model III	Probability calculated from model II + postoperative information (residual disease, intraoperative and postoperative complications)

Data analysis and machine learning approach

- Data will be described using mean, median, standard deviation, range and percentage depending on variable type and distribution.
- Missing data will be treated using MICE (Multiple Imputation by Chained Equations) and KNN (k-Nearest Neighbors). Both approaches will be tested on a virtual missing data. Virtual missing data will be created by deleting 30% of data from variables that show most missing data and accuracy of prediction of missing values will be compared between both methods to decide which method works the best. The selected method will be applied to the whole database.
- Data will be utilized as Xi and Yi where Xi presents input (features) and Yi presents dependent variables (outcomes). Four algorithms will be tested for the highest accuracy (logistic regression, Support vector machine, Xgboost, and random forest). A train-test split will be applied to data using 0.8:0.2 ratio. Model performance will be evaluated using accuracy, F1 score, precision and recall scores. Model training will be assessed using plotted learning curves which will be validated using K-fold cross-validation to test for over-fitting and plateauing of learning process.
- Three models will be designed for each outcome including overall survival and disease-free survival at 1, 3 and 5 years. Model I will include patient characteristics and preoperatively determined disease characteristics. This model aims at determining probability of survival at the time of disease diagnosis. Calculated probability from model I will be included as a feature in model II, in addition to planned surgery and adjuvant therapy. The aim of this model is to determine prognosis per proposed management plan(s) in priori. Model III will include probability from model II in addition to features determined after surgery e.g., residual disease and treatment complications (table).
- Calculated probability will be plotted against each outcome using receiver operating characteristic (ROC) curve to test diagnostic performance of these models.
- Machine learning models will be created using using python 3.8®, through Anconda 3® interface. A software will be designed to facilitate the use of these models in clinical practice.

References

1. Bell, D.W. and L.H. Ellenson, Molecular Genetics of Endometrial Carcinoma. Annu Rev Pathol, 2019. 14: p. 339-367.
2. Lax, S.F., [New features in the 2014 WHO classification of uterine neoplasms]. Pathologe, 2016. 37(6): p. 500-511.
3. Hamilton, C.A., et al., Endometrial cancer: A society of gynecologic oncology evidence-based review and recommendations. Gynecol Oncol, 2021.
4. Soja, M., et al., Analysis of the results of invasive diagnostic procedures in patients referred to gynecologic department due to abnormal uterine bleeding. Prz Menopauzalny, 2020. 19(4): p. 155-159.
5. Bray, F., et al., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin, 2018. 68(6): p. 394-424.
6. Ignatov, A. and O. Ortmann, Endocrine Risk Factors of Endometrial Cancer: Polycystic Ovary Syndrome, Oral Contraceptives, Infertility, Tamoxifen. Cancers (Basel), 2020. 12(7).
7. Raglan, O., et al., Risk factors for endometrial cancer: An umbrella review of the literature. Int J Cancer, 2019. 145(7): p. 1719-1730.
8. Kruse, A.J., et al., Vaginal hysterectomy with or without bilateral salpingo-oophorectomy may be an alternative treatment for endometrial cancer patients with medical co-morbidities precluding standard surgical procedures: a systematic review. Int J Gynecol Cancer, 2019.
9. Murali, R., et al., Evolving Roles of Histologic Evaluation and Molecular/Genomic Profiling in the Management of Endometrial Cancer. J Natl Compr Canc Netw, 2018. 16(2): p. 201-209.
10. Petousis, S., et al., Combined pelvic and para-aortic is superior to only pelvic lymphadenectomy in intermediate and high-risk endometrial cancer: a systematic review and meta-analysis. Arch Gynecol Obstet, 2020. 302(1): p. 249-263.
11. Mariani, A., et al., Low-risk corpus cancer: is lymphadenectomy or radiotherapy necessary? Am J Obstet Gynecol, 2000. 182(6): p. 1506-19.
12. Reijnen, C., et al., Preoperative risk stratification in endometrial cancer (ENDORISK) by a Bayesian network model: A development and validation study. PLoS Med, 2020. 17(5): p. e1003111.