Published in final edited form as:

Gynecol Oncol. 2016 July ; 142(1): 150–157. doi:10.1016/j.ygyno.2016.05.012.

Stratification of endometrioid endometrial cancer patients into risk levels using somatic mutations

Donghai Dai^a, Kristina W. Thiel^a, Erin A. Salinas^a, Michael J. Goodheart^{a,b}, Kimberly K. Leslie^{a,b}, and Jesus Gonzalez Bosquet^{a,*}

^aDepartment of Obstetrics and Gynecology, University of Iowa, Iowa City, IA, United States

^bHolden Comprehensive Cancer Center, University of Iowa, Iowa City, IA, United States

Abstract

Objective—Patients with endometrioid endometrial cancer are stratified as high risk and low risk for extrauterine disease by surgical staging. Since patients with low-grade, minimally invasive disease do not benefit from comprehensive staging, pre-surgery stratification into a risk category may prevent unnecessary surgical staging in low risk patients. Our objective was to develop a predictive model to identify risk levels using somatic mutations that could be used preoperatively.

Methods—We classified endometrioid endometrial cancer patients in The Cancer Genome Atlas (TCGA) dataset into high risk and low risk categories: high risk patients presented with stage II, III or IV disease or stage I with high-intermediate risk features, whereas low risk patients consisted of the remaining stage I patients with either no myometrial invasion or low-intermediate risk features. Three strategies were used to build the prediction model: 1) mutational status for each gene; 2) number of somatic mutations for each gene; and 3) variant allele frequencies for each somatic mutation for each gene.

Results—Each prediction strategy had a good performance, with an area under the curve (or AUC) between 61% and 80%. Analysis of variant allele frequency produced a superior prediction model for risk levels of endometrial cancer as compared to the other two strategies, with an AUC = 91%. Lasso and Ridge methods identified 53 mutations that together had the highest predictability for high risk endometrioid endometrial cancer.

Conclusions—This prediction model will assist future retrospective and prospective studies to categorize endometrial cancer patients into high risk and low risk in the preoperative setting.

GRAPHICAL ABSTRACT

*Corresponding author at: Department of Obstetrics & Gynecology, University of Iowa Hospitals and Clinics, 200 Hawkins Dr., Iowa City, IA 52242, United States. jesus-gonzalezbosquet@uiowa.edu (J. Gonzalez Bosquet).

Disclosure of potential conflict of interest

Drs. Donghai Dai and Kristina W. Thiel are co-owners of Immortagen, L.L.C. All other authors declare no competing interests.

Authors' contributions

DD and JGB, contributed in the data collection, conception and study design, data and computational analysis, interpretation of results and manuscript writing. KWT and KKL, contributed in the conception and study design, interpretation of results and manuscript writing and editing. AES and MJG contributed in the data collection, analysis interpretation and manuscript editing.

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.ygyno.2016.05.012>.

A prediction model with variant allele frequencies of 53 somatic gene mutations from TCGA (A), located throughout the genome (B), stratified endometrioid endometrial cancer patients in risk levels with an AUC of 91% (C).

Keywords

Endometrial cancer; Somatic mutations; Prediction model; Risk levels; Individualized treatment

1. Introduction

Endometrial cancer is the most common gynecological cancer, with over 54,000 new cases predicted for 2015 [1]. >70% of all endometrial tumors are at early stage (stage I) at the time of diagnosis [2]. Type I, or endometrioid endometrial (EEC), is the most common histologic type of endometrial cancer, and most are low grade and confined to the uterus. Surgical staging, including removing the uterus, cervix, adnexa, and pelvic and para-aortic lymph node tissues as well as obtaining pelvic washings, is the standard of care to accurately stage and assign patients into high risk and low risk categories [2-5], which can inform subsequent treatment [6]. Low risk patients have been defined as those with very low risk of having extrauterine disease and thus not requiring further treatment beyond a simple hysterectomy and removal of adnexa, preferably through a minimally invasive approach [7-9]. In comparison, high risk patients are at risk of having extrauterine disease, present worse prognosis and will likely require adjuvant treatment after surgery, including brachytherapy, external beam radiation or systemic chemotherapy.

Recent clinical trials have demonstrated that low risk patients with low-grade, minimally invasive disease do not clearly benefit from comprehensive staging. In fact, such patients will have excellent outcomes even with limited surgical intervention, hysterectomy and removal of the adnexa, with no further adjuvant therapy [4,10]. Moreover, the complex surgical procedures used for staging also increase complication rates and overall cost of care [11]. Thus, several groups have tried to identify predictors of disseminated disease to limit comprehensive surgical staging to those patients that will benefit (i.e., high risk patients) [12-15]. The Gynecologic Oncology Group (GOG) and other groups determined that histological grade and depth of myometrial invasion are associated with extension of disease outside the uterus [5,13,15]. However, the assessment of these variables may be limited by inaccurate and unreliable analysis performed on frozen specimens during the surgical

procedure [16], highlighting the need for alternative approaches to stratify patients into risk categories.

Prediction models designed to assess high risk endometrial cancer have been constructed with information from surgical specimens [5, 15,17,18]. These models usually are excellent at detecting patients at low risk, with negative predictive values (NPV) close to 100%. However, they are not able to accurately detect patients as high risk, and their positive predictive value (PPV) is at most fair, around 20% in EEC [17,18]. Prediction models of risk using preoperative data have the same limitations [16,19]. In general, with these algorithms, it is necessary to perform 4 to 8 lymphadenectomies to find one patient with positive lymph nodes [19], and this surgery is not without extra costs and complications. Thus, to date, there is no preoperative predictive model that accurately identifies women with high risk EEC [16].

Our objective in this study was to create a test for risk using somatic mutations from whole exome next generation sequencing (NGS) that could be used preoperatively. Prediction models based on variant allele frequencies for somatic mutations were able to discriminate low versus high risk EEC with a mean area under the ROC (receiver operator characteristic) curve, or AUC, of 91%.

2. Materials and methods

2.1. Patients and data collection

Patients were selected from the Cancer Genome Atlas (TCGA) database of endometrial cancer. Patients with serous histology and other Type II endometrial cancer were excluded. Of those patients with Type I endometrial cancer or EEC, we only included patients who had undergone whole exome next generation sequencing with a somatic mutation report from TCGA (n = 190) [20]. Assuming that patients with myometrial invasion <50% (2009 FIGO stage IA) and histological grade 1 or 2 rarely would need lymph node assessment, 182 out of 190 patients that considered fully staged, or 96%. Mutations from these TCGA patients were downloaded from two different sources:

1. Level 2 and 3 of mutation analysis from the exomes of 190 ECC tumors were sequenced on Illumina GAIIX or HiSeq 2000 platforms (Illumina Inc., San Diego, CA). Somatic single variants and indels were called using software described elsewhere and were filtered for potential false positives [20]. The final list of somatic mutations in EEC is available online (<https://tcga-data.nci.nih.gov/tcga/>). A total of 177,057 somatic mutations were identified in the targeted exons of the 19,552 genes analyzed in EEC samples.
2. Original BAM files for each endometrial cancer patient were downloaded with permission from NCI from Cancer Genomics Hub (<https://cghub.ucsc.edu>). SAMtools [21] was then used to pile reads against Human Reference Genome 19. All reads for every synonymous (silent) and nonsynonymous mutation site were obtained. The Bcftools package in SAMtools was used to call variant sites from this subset of loci. Only sites for which one allele matched the reference genome were used in this analysis. Next, the number of reads for the reference

and the strongest non-reference allele were calculated by summing the number of forward and reverse paired-end reads for each site. The strongest non-reference allele helped alleviate concerns of sequencing errors because each site required multiple reads per allele (>3) before being considered as a variant site. As an additional layer of stringency, only reads with Phred quality score of >30 (i.e. 99.9% accurate) were used in this study. The variant allele frequency (VAF) was calculated by the number of variant allele reads divided by the total number of reads over a genome locus. Thus, every patient was presented with a variant allele frequency for each somatic mutation.

2.2. Classification of EEC risk

Classification of EEC risk was based on the results and criteria from GOG 33 study, GOG 99 clinical trial and modified in the PORTEC trials [3,5,15,22]. “High risk” patients were defined as those at risk of having extrauterine disease and most likely needing any type of adjuvant treatment after surgery. Specifically, all patients presenting with stage II, III and IV as defined by 2009 FIGO classification (and sanctioned in 2014) [23], and patients with initial stage I and high-intermediate risk features by GOG 99 criteria [5] were classified as high risk. High-intermediate features of stage I included three risks factors: 2 or 3 tumor grade, presence of lymphovascular invasion, and outer-third myometrial invasion, with the following criteria: 1) at least 70 years of age with only one of the risk factors, 2) at least 50 years of age with any two of the other risk factors, or 3) any age with all three of the other risk factors. High risk patients are also at higher risk for disease recurrence and needing adjuvant treatment [6,24]. “Low risk” patients were the remaining stage I patients, either with no myometrial invasion or low-intermediate risk features by GOG 99 criteria [5]. There were 62 high risk and 128 low risk patients available for the study. A summary of clinical and pathological characteristics of both risks groups are shown in Supplemental Table S1. In the multivariable analysis of these features, there were significant differences between those variables that defined low and high risks groups: age, stage, myometrial invasion and histological grade. The stratification by risk was also independently associated with survival in TCGA patients with EEC (Supplementary Fig. S2). Five-year overall survival for low risk patients was 93% versus 82% for high risk patients.

2.3. Prediction model construction

Three strategies were used to build the prediction model: 1) mutational status; 2) the number of somatic mutations for each gene abstracted from the Level 2 dataset from TCGA; 3) variant allele frequency for each somatic mutation derived from reanalysis of BAM files downloaded from TCGA.

2.3.1. Strategy #1—In the *first strategy*, the mutational status for each gene in individual patients was reported as level 3 exome sequencing data in TCGA portal. A design matrix was created with patients as rows and all mutated genes as column and mutational status represented by 0 (no mutation) and 1 (mutated). The risk classification of high and low was represented as 1 and 0, respectively. A ten-fold cross-validation was used for each Lasso analysis to assess how accurately the groups (high vs low risk) were predicted and to avoid over-fitting. For all the prediction methods, cross-validation was performed to limit over-

fitting of the prediction model. We used k -fold cross-validation where the samples are randomly partitioned into approximately k equal size groups. One of the k subsets is omitted and the classifier model developed using a training set consisting of samples in the union of the other $k - 1$ subsets. This is done k times, omitting each of the k subsets one at a time as validation of the initial model. When the number of samples is large, a 10-fold cross-validation is typically used and has been suggested to provide a more precise estimate [25]. The analysis was repeated 10 times and the performance of the prediction model was reported as mean area under the curve (AUC) value. A flow diagram delineating the strategies and methods used for the analysis, including this strategy, are detailed in Fig. 1.

2.3.2. Strategy #2—The *second strategy* used the number of mutations per gene in EEC samples for prediction of high and low risk for adjuvant treatment in EEC (Fig. 1). The ‘*Classification for MicroArrays*’ (CMA) tool was applied to TCGA mutation dataset. CMA is a statistical tool designed to construct and evaluate classifiers derived from microarray experiments using a large number of usual methods [26] and utilizes R environment for statistical computing (www.r-project.org) [27], and Bioconductor packages as open source software for bioinformatics (bioconductor.org). Initially, a logistic regression analysis was used to determine which genes included in the EEC TCGA mutation database had a significantly different number of mutations between patients with low versus high risk, with a p -value < 0.05 . Next, all these genes (411 in total) were used to compute the prediction model of risk for adjuvant treatment in EEC, referred as 411-gene *complete prediction model*. To assess prediction accuracy and to avoid over-fitting, a 10-fold cross-validation was performed (internal validation of the classifier) [25]. The prediction performance was computed with corrections for TCGA batch-effect. No other variables were associated independently with risk levels but those used for its construction include the following: age, stage, myometrial invasion, and tumor grade (Supplementary Table S1). Different methods available in the CMA package were used to create the model. The performance of this model was compared in terms of sensitivity, specificity, misclassification rate and AUC (ROC - receiver operator characteristic curve). For each of the AUC measurements, we also computed a 95% confidence interval (CI), which estimates with a 95% probability where the true AUC value is contained. CI of AUC was also used to compare different models and different methods of classification. To illustrate the performance of the predictor in classifying EEC risk, a ROC curve was generated.

To further optimize the prediction model for risks levels in EEC, the number of mutated genes included in the model were narrowed down from the initial 411 mutated genes associated with high or low risk, to those most informative for the prediction of risk [28,29]. The intention of the variable selection process is to optimize the prediction model by selecting those variables that are more informative in the prediction process, discard those that do not add anything to the model (potential for misclassification or confusion), and simplify the model for latter clinical application. The selection process was performed through 11 different methods that ranked genes from the most relevant to the least relevant in creating the model: two-sample t -test; Welch modification of the t -test; Wilcoxon rank sum test; F-test; Kruskal-Wallis test; “moderated” t and F test, respectively, using the package ‘limma’ in R statistics; one-step Recursive Feature Elimination (RFE) in

combination with the linear support vector machines (SVM); random forest variable importance measure; least absolute shrinkage and selection operator (or Lasso); the a regularized regression method or elastic net; component-wise boosting; and ad-hoc “Golub” criterion [26]. Using the gene selection tool of the software package, each gene was ranked depending on its relative importance in prediction models. These genes were ordered based on their rank (or relative ‘weight’) in the prediction process, and the prediction model analysis was applied by including only those genes that had been ranked at least once by each method, or at least 11 times in total. This model which included only the selected and more informative genes was termed the “simplified prediction model.”

2.3.3. Strategy #3—For the *third analysis strategy*, the somatic mutation data was downloaded from TCGA data portal and a predictor matrix was constructed with patients as rows and the genes that carry at least one somatic mutation as columns (Fig. 1). Instead of using 0 or 1 to indicate if there was a somatic mutation for a specific gene in an individual, variant allele frequency was used in the cell for the predictor matrix. This matrix was used to train the lasso and ridge methods using the glmnet package in R project for Statistic Computing (<http://www.r-project.org/>). Tenfold cross-validation was adopted to train prediction for high and low risk patients as defined above. The predicted values were entered into the ROCR package to evaluate performance and AUC values. The same process was repeated 10 times and then the average AUC was reported.

2.4. Software

The ‘*Classification for MicroArrays*’ (CMA) utilized for the prediction analysis is described above. Logistic regression, Cox proportional Hazard ratio, survival curves, were performed using R statistical package for statistical computing and graphics as background and Bioconductor packages as open source software for bioinformatics (bioconductor.org). Differential gene expression was performed using Biometric Research Branch (BRB) ArrayTools, an integrated package for the visualization and statistical analysis that utilizes Excel (Microsoft, Redmond, WA) as the front end, and with tools developed in the R statistical system. BRB-ArrayTools were developed by Dr. Richard Simon and the BRB-ArrayTools development team.

3. Results

3.1. Strategy #1: prediction models using mutational status for each gene in EEC patients

Since the goal of this study is to develop a better pre-treatment predictive model to stratify patients into high risk vs. low risk, we first established the performance of the current prediction model constructed using the clinical variables of age and tumor grade (Table 1). These variables were specifically chosen because previous reports have established that age and tumor grade are independently associated with risk level [3-6,13], and these are the only independently associated clinical variables that are available preoperatively. We found that clinical variables alone resulted in an AUC range of 70–82%, depending upon the different method used for analysis (Table 1). Next, the mutational status for each gene in individual patients reported from TCGA level 3 exome sequencing was used for the prediction analysis. After 10 replications of the Lasso analysis with a 10-fold cross-validation, the

mean AUC value was 77.6% (Supplementary Fig. S3), with a range from 68.7% to 80.9%. Fourteen genes were identified as predictive genes to determine the level of risk for EEC: *DHX35*, *FGF12*, *FYN*, *HOXA13*, *HS1BP3*, *PAPD4*, *PTDSS1*, *RIMKLA*, *SENP5*, *SH3BP4*, *SLC2A4*, *SLC3A2*, *THBS2* and *TNKS1BP1*. Of note, the prediction performance of this analysis was not better than the performance of the prediction model constructed only with clinical variables (Table 1).

3.2. Strategy #2: prediction models using number of mutations per gene in EEC patients

3.2.1. Analysis of genes with different number of mutations between HR and LR patients—Initially, we performed a logistic regression analysis to determine which of the 19,552 genes included in the EEC TCGA mutation database had a significantly different number of mutations between low and high risk patients, with a p-value < 0.05. A total of 411 genes had different somatic mutations between the two groups (Supplementary Table S5), and all of these genes were used to compute the prediction model of risk for adjuvant treatment in EEC, referred to as the 411-gene *complete prediction model*. The number of POLE mutations was not significantly different between high and low risk patients: 16.1% (or 10/62) and 14.8% (or 19/128) respectively, consistent with a previous report [30].

3.2.2. 411-gene complete prediction model—The *complete prediction model* included all 411 genes with different number of mutations between high and low risk patients (Supplementary Table S5). The performance of the prediction model computed with the CMA software suite was measured in terms of AUC for eight different methods and with their respective CI, with results that ranged from 44 to 63% (Fig. 2A). However, the performance of this model was not better than the model using only clinical variables alone (Table 1 and Fig. 2C).

3.2.3. Simplified prediction model with most informative genes—The selection process with 11 different methods identified 109 different genes, out of the significant 411, to be relevant, or with at least with one ‘hit’, in the prediction model. Only 35 of those 109 genes had 11 or more ‘hits,’ with on average one ‘hit’ per method. The most relevant mutated genes for the construction of the model are *ADAMTS17*, *DVL2*, *CYP39A1*, *WDR3*, *COLQ*, *PGBD5*, *SBF1*, *ENSG00000171101*, *DDX49*, *RUSC2*, *PIAS4*, *PRRC2A*, *RPS6KA2*, *GNAO1*, *BARHL2*, *GAL3ST1*, *OAT*, *RAG2*, *TECTA*, *ADAM22*, *GPRIN2*, *BRPF3*, *CELSR3*, *CTGF*, *KIAA0319*, *NDUFAF1*, *NTNG1*, *CHRNA2*, *SLC13A5*, *TAAR6*, *ABCF3*, *KCTD18*, *MAP7D1*, *MPRIIP*, and *ST3GAL1*. A new prediction of risk for adjuvant treatment was performed with these 35 genes (termed the “simplified model”), which also accounts for batch-effect and uses cross-validation (Fig. 2B). By selecting only those genes that were more informative in the prediction model and eliminating those that had little or no influence, the performance of the signature in terms of AUC increased by 5–15%, though this model was not superior to the clinical prediction model (Table 1 and Fig. 2C).

3.3. Strategy #3: prediction models based on variant allele frequency for each somatic mutation in EEC samples

Among all genes with somatic mutations in EEC, the variant allele frequency provided a better prediction of high risk patients. Specifically, a 10-fold cross-validation was used for

each Lasso analysis, and the mean of all ten AUC analyses, as a measure of prediction performance, was 90.8%, with a range from 87.8% to 92.7% (Fig. 3). Importantly, this model out-performed the clinical model, with ranges above of those of the clinical model, Table 1, Fig. 2). Fifty-three genes were identified as predictive genes to determine the level of risk for EEC (Fig. 4A) The relative chromosomal position of these selected 53 mutated genes and their differential gene expression between high and low risk are represented in a circular layout with matrix depiction (Fig. 4B). Adding clinical variants to these 53 genes, did not improved the prediction model.

4. Discussion

More than half of patients with EEC are considered to be low risk, with disease limited to the uterus, and the majority will not require further treatment after surgery [24]. These low risk patients most likely would benefit from a simple hysterectomy and removal of adnexa without further surgical staging, preferably through a minimally invasive approach [7-9]. To date, however, current prediction models of risk are constructed with information from surgical specimens [5,15,17,18], and there is no preoperative predictive model that accurately identifies women with low or high risk disease [16,19]. The result is that between four and eight unnecessary surgical staging procedures are performed to identify one patient with extrauterine disease [16], with additional costs and surgical complications.

Previous studies have tried to classify endometrial cancer based on molecular characteristics found on TCGA analyses [20,30,31]. The goal of these studies was to create prediction models that would classify patients into different prognostic groups. While this is very valuable to the clinician and patients to assess the severity of the disease, prognosticators have limited immediate clinical application. In other words, it may take time to translate this prognosis classification into changes in treatment. Our study design was aimed to have immediate effect on treatment decisions. Our objective was to create a prediction model that would distinguish low risk patients that were unlikely to benefit from further treatment from high risk patients that would need more aggressive surgical approach, staging, and most likely some type of adjuvant treatment. Another goal for this predictive model for risk was the potential to be used in the preoperative setting. This is possible because, for endometrial cancer, we already have very good prognosticators that have been validated retrospective and prospectively [5,13, 15-18]. In fact, the prediction model constructed in our study only with some of these clinical variables (age and histological grade) had a good performance, with AUCs ranging from 75 to 82%. The first two strategies that we report using TCGA somatic mutation data were inferior to the clinical prediction model. These strategies included: 1) the status of the mutation, or the presence or absence of a mutation in a determined gene (strategy #1); 2) the number of mutations per gene for each patient (strategy #2). Only the strategy that takes into account mutation allele frequencies (strategy #3) was considerably superior to the clinical model, with a mean AUC of 91% (range of 88 to 93%).

Having a preoperative tool that could stratify patients accordingly to risk has the potential to dramatically change the clinical management of patients with EEC. Patients could be given a better prognosis of their disease before starting any treatment, with better information about possible surgical options, like minimally invasive surgery with overnight stay in the hospital

versus more aggressive surgical approach and potential longer stay and more complications. Physicians could schedule surgical procedures more accurately without adding potential time for surgical staging that may or may not happen, and would be able to streamline low impact treatment for low risk patients. From an administrative and payer point of view, having a better understanding of the amount of care, surgical effort, hospitalization and postsurgical treatment that have to be dedicated to certain patients would help to standardize care and better distribute resources, potentially lowering health care costs and improving patient outcomes.

Cancer is a genetic disease, developed through sequential accumulation of somatic mutations [32]. The progression of cancer, from a normal cell to atypical cell to advanced stages of cancer, is also an evolution process where the occurrence of each driver mutation confers the host cells with a growth advantage [32,33]. Clonal evolution among cancer subpopulations leads to the development of intratumoral heterogeneity [34,35]. We recognize that intratumoral heterogeneity and tumor sampling (tumor purity) are important concerns. It is generally acceptable to use variant allele frequency as a way to represent the proportion of cancer cells that harbor individual mutations, as we used herein. However, currently the best available methodology to fully understand the heterogeneous distribution of mutations (“intratumoral heterogeneity”) and determine the “cancer cell fraction” for each mutation is to perform NGS of subclones with simultaneous consideration of copy number changes at the locus of the specific mutation. This laborious and expensive approach is not feasible for clinical application. Regarding tumor purity, TCGA implemented rigorous specimen quality control to ensure a high purity of the cancer cell component of the tumor for NGS. Their purity levels are much higher than most published data. However, careful quality control for endometrial tumor biopsies and preparation of samples for NGS will be essential for the future clinical application of our model to stratify patients into risk categories.

TCGA project has characterized somatic mutations in thousands of human tumors across many cancer types using NGS technology [36]. NGS is a powerful technology that can detect reliably targeted mutations from small amounts of initial DNA, including fresh frozen biopsies, biopsies embedded in paraffin, or even circulating tumor DNA (ctDNA or liquid biopsy) [37,38]. However, somatic mutations downloadable at the TCGA portal only report mutated genes in individual samples without detailed information regarding if the mutation is present in all cancer cells or in only a portion of the tumor. Variant allele frequency, calculated by the ratio of the number of reads harboring a mutation and the number of total reads that cover the position, reveals the prevalence of a mutated allele and could be used as a valuable measurement of intra-tumoral heterogeneity at the time of tumor specimen collection. Of note, the model that combined the weight of the predictive genes and variant allele frequency values for the prediction of high and low risk patients achieved the highest accuracy with a mean AUC of 91%. The prediction based on variant allele frequency for each mutation appears to be better than that computed by the simple number of somatic mutations by gene, suggesting that intra-tumoral heterogeneity embedded in varying variant allele frequencies for somatic mutations could be useful to predict disease aggressiveness. Thus, deep sequencing coverage of the targeted genes in combination with determination of variant allele frequencies for those somatic mutations, as performed in our study, may help to account for intratumoral heterogeneity [39,40].

This study was performed using one of the most comprehensive databases of somatic mutations from EEC specimens publicly available (TCGA) [20], and the prediction models were computed with a broad spectrum of methods to test their performance [26]. Despite the internal cross-validation during prediction modeling, a limitation of this study is that external validation in independent databases was not possible due to the lack of publicly available datasets with large sample sizes of EEC with complete somatic mutation analysis. Once the current algorithm has been validated in independent databases and refined, it could be combined with other strategies already published [5,12-15] to enhance the individual management of patients with EEC.

Our study was performed with tumors dissected from large surgical specimens at the TCGA. However, retrospective studies have established that molecular parameters are highly concordant between diagnostic preoperative biopsies and surgical specimens [31]. Therefore, molecular-based prediction models from presurgical biopsy specimens may be consistent with prediction models created with surgical specimens, despite the potential sampling error [31]. To this extent there are prospective initiatives to determine the validity of molecular classification in diagnostic endometrial samples and their accuracy [30]. We anticipate that variables available in the preoperative biopsy, such as tumor grade, would maximize the performance of the prediction model.

The future clinical deployment of the prediction model to stratify patients into risk categories preoperatively is high, and we envision the following process to become standard-of-care for all new diagnoses of endometrial cancer. Once a biopsy has been obtained from a symptomatic patient and the diagnosis of endometrioid endometrial cancer established, the tumor specimen would be subjected to sequencing for the 53 genes found to be predictive of levels of risk in the strategy #3. Sequencing the tumor specimen is not a significant technical hurdle since numerous studies have provided compelling evidence that DNA sequencing can be accomplished using small DNA quantities (as low as 1 ng), even from paraffin-embedded tissue [37], and with a rapid turn-around time that will not extend the average time from diagnosis to surgery beyond the current standard of three weeks. Sequencing data would be used to determine allele frequencies of the mutated genes. Next, we will apply an algorithm that takes into account this mutational allele frequency of the 53 genes and their respective weights, or relative importance in the prediction process. This will create a numeric score for each individual patient. Additional studies are necessary to fine-tuning of the prediction model, including setting a threshold value for this numeric score that can accurately discriminate between high and low risk patients with a 95% CI. Based upon stratification of the patient into low/high risk classification, the gynecologic oncologist will choose whether the patient will receive minimally invasive surgery vs. more aggressive surgical staging and adjuvant therapy.

In summary, we have identified a prediction model including gene mutations as a surrogate for tumor heterogeneity that can stratify EEC patients into low and high risk. This classifier may impact not only the invasiveness of surgery to remove the primary tumor, but also guide the use of adjuvant treatment following surgery. Accordingly, better risk stratification and treatment selection could lead to substantial decreases in treatment complications and optimization of resource allocation and implementation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support

Supported in part by the NIH grant CAR0199908 in endometrial cancer, and the Basic research fund from the Department of Obstetrics & Gynecology University of Iowa.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. *Ca-a Cancer Journal for Clinicians*. 2015; 65:5–29. [PubMed: 25559415]
2. Creasman WT, Odicino F, Maisonneuve P, Quinn MA, Beller U, Benedet JL, et al. Carcinoma of the corpus uteri. FIGO 26th Annual Report on the Results of Treatment in Gynecological Cancer. *Int J Gynaecol Obstet*. 2006; 95(Suppl 1):S105–S143. [PubMed: 17161155]
3. Creutzberg CL, van Putten WL, Koper PC, Lybeert ML, Jobsen JJ, Warlam-Rodenhuis CC, et al. Surgery and postoperative radiotherapy versus surgery alone for patients with stage-1 endometrial carcinoma: multicentre randomised trial. PORTEC Study Group. *Post Operative Radiation Therapy in Endometrial Carcinoma*. *Lancet*. 2000; 355:1404–1411. [PubMed: 10791524]
4. Kitchener H, Swart AM, Qian Q, Amos C, Parmar MK. ASTEC study group. Efficacy of systematic pelvic lymphadenectomy in endometrial cancer (MRC ASTEC trial): a randomised study. *Lancet*. 2009; 373:125–136. [PubMed: 19070889]
5. Keys HM, Roberts JA, Brunetto VL, Zaino RJ, Spirtos NM, Bloss JD, et al. A phase III trial of surgery with or without adjunctive external pelvic radiation therapy in intermediate risk endometrial adenocarcinoma: a Gynecologic Oncology Group study. *Gynecol Oncol*. 2004; 92:744–751. [PubMed: 14984936]
6. Endometrial cancer. Practice Bulletin No 149. *Obstet Gynecol*. 2015; 125:1006–1026. [PubMed: 25798986]
7. Walker JL, Piedmonte MR, Spirtos NM, Eisenkop SM, Schlaerth JB, Mannel RS, et al. Laparoscopy compared with laparotomy for comprehensive surgical staging of uterine cancer: Gynecologic Oncology Group Study LAP2. *J Clin Oncol*. 2009; 27:5331–5336. [PubMed: 19805679]
8. Kornblith AB, Huang HQ, Walker JL, Spirtos NM, Rotmensch J, Cella D. Quality of life of patients with endometrial cancer undergoing laparoscopic international federation of gynecology and obstetrics staging compared with laparotomy: a Gynecologic Oncology Group study. *J Clin Oncol*. 2009; 27:5337–5342. [PubMed: 19805678]
9. Janda M, Gebiski V, Brand A, Hogg R, Jobling TW, Land R, et al. Quality of life after total laparoscopic hysterectomy versus total abdominal hysterectomy for stage I endometrial cancer (LACE): a randomised trial. *Lancet Oncol*. 2010; 11:772–780. [PubMed: 20638899]
10. Benedetti Panici P, Basile S, Maneschi F, Alberto Lissoni A, Signorelli M, Scambia G, et al. Systematic pelvic lymphadenectomy vs. no lymphadenectomy in early-stage endometrial carcinoma: randomized clinical trial. *J Natl Cancer Inst*. 2008; 100:1707–1716. [PubMed: 19033573]
11. Dowdy SC, Borah BJ, Bakkum-Gamez JN, Weaver AL, McGree ME, Haas LR, et al. Prospective assessment of survival, morbidity, and cost associated with lymphadenectomy in low-risk endometrial cancer. *Gynecol Oncol*. 2012; 127:5–10. [PubMed: 22771890]
12. Todo Y, Okamoto K, Hayashi M, Minobe S, Nomura E, Hareyama H, et al. A validation study of a scoring system to estimate the risk of lymph node metastasis for patients with endometrial cancer for tailoring the indication of lymphadenectomy. *Gynecol Oncol*. 2007; 104:623–628. [PubMed: 17097721]

13. Mariani A, Webb MJ, Keeney GL, Haddock MG, Calori G, Podratz KC. Low-risk corpus cancer: is lymphadenectomy or radiotherapy necessary? *Am J Obstet Gynecol.* 2000; 182:1506–1519. [PubMed: 10871473]
14. Hidaka T, Nakashima A, Shima T, Hasegawa T, Saito S. Systemic lymphadenectomy cannot be recommended for low-risk corpus cancer. *Obstet Gynecol Int.* 2010; 2010:490219. [PubMed: 20168975]
15. Creasman WT, Morrow CP, Bundy BN, Homesley HD, Graham JE, Heller PB. Surgical pathologic spread patterns of endometrial cancer, A Gynecologic Oncology Group Study. *Cancer.* 1987; 60:2035–2041. [PubMed: 3652025]
16. AlHilli MM, Mariani A. Preoperative selection of endometrial cancer patients at low risk for lymph node metastases: useful criteria for enrollment in clinical trials. *J Gynecol Oncol.* 2014; 25:267–269. [PubMed: 25310031]
17. Mariani A, Dowdy SC, Cliby WA, Gostout BS, Jones MB, Wilson TO, et al. Prospective assessment of lymphatic dissemination in endometrial cancer: a paradigm shift in surgical staging. *Gynecol Oncol.* 2008; 109:11–18. [PubMed: 18304622]
18. Convery PA, Cantrell LA, Di Santo N, Broadwater G, Modesitt SC, Secord AA, et al. Retrospective review of an intraoperative algorithm to predict lymph node metastasis in low-grade endometrial adenocarcinoma. *Gynecol Oncol.* 2011; 123:65–70. [PubMed: 21742369]
19. Mitamura T, Watari H, Todo Y, Kato T, Konno Y, Hosaka M, et al. Lymphadenectomy can be omitted for low-risk endometrial cancer based on preoperative assessments. *J Gynecol Oncol.* 2014; 25:301–305. [PubMed: 25142623]
20. Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, et al. Network CGAR. Integrated genomic characterization of endometrial carcinoma. *Nature.* 2013; 497:67–73. [PubMed: 23636398]
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009; 25:2078–2079. [PubMed: 19505943]
22. Nout RA, Smit VT, Putter H, Jurgensliemk-Schulz IM, Jobsen JJ, Lutgens LC, et al. Vaginal brachytherapy versus pelvic external beam radiotherapy for patients with endometrial cancer of high-intermediate risk (PORTEC-2): an open-label, non-inferiority, randomised trial. *Lancet.* 2010; 375:816–823. [PubMed: 20206777]
23. FIGO Committee on Gynecologic Oncology. FIGO staging for carcinoma of the vulva, cervix, and corpus uteri. *Int J Gynaecol Obstet.* 2014; 125:97–8. [PubMed: 24630859]
24. Morice P, Leary A, Creutzberg C, Abu-Rustum N, Darai E. Endometrial cancer. *Lancet.* 2015
25. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol.* 2005; 23:7332–7341. [PubMed: 16145063]
26. Slawski M, Daumer M, Boulesteix AL. CMA: a comprehensive bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics.* 2008; 9:439. [PubMed: 18925941]
27. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2014. at <http://www.R-project.org/>
28. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst.* 2003; 95:14–18. [PubMed: 12509396]
29. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010; 21:128–138. [PubMed: 20010215]
30. Talhouk A, McConechy MK, Leung S, Li-Chang HH, Kwon JS, Melnyk N, et al. A clinically applicable molecular-based classification for endometrial cancers. *Br J Cancer.* 2015; 113:299–310. [PubMed: 26172027]
31. Stelloo E, Nout RA, Naves LC, Ter Haar NT, Creutzberg CL, Smit VT, et al. High concordance of molecular tumor alterations between pre-operative curettage and hysterectomy specimens in patients with endometrial carcinoma. *Gynecol Oncol.* 2014; 133:197–204. [PubMed: 24556061]
32. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell.* 1990; 61:759–767. [PubMed: 2188735]

33. Jones S, Chen WD, Parmigiani G, Diehl F, Beerenwinkel N, Antal T, et al. Comparative lesion sequencing provides insights into tumor evolution. *Proc Natl Acad Sci U S A*. 2008; 105:4283–4288. [PubMed: 18337506]
34. Nowell PC. The clonal evolution of tumor cell populations. *Science*. 1976; 194:23–28. [PubMed: 959840]
35. Fearon ER, Hamilton SR, Vogelstein B. Clonal analysis of human colorectal tumors. *Science*. 1987; 238:193–197. [PubMed: 2889267]
36. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. *Nature*. 2013; 502:333–339. [PubMed: 24132290]
37. Murtaza M, Dawson SJ, Pogrebniak K, Rueda OM, Provenzano E, Grant J, et al. Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. *Nat Commun*. 2015; 6:8760. [PubMed: 26530965]
38. de Leng WW, Gadellaa-van Hooijdonk CG, Barendregt-Smouter FA, Koudijs MJ, Nijman I, Hinrichs JW, et al. Targeted next generation sequencing as a reliable diagnostic assay for the detection of somatic mutations in tumours using minimal DNA amounts from formalin fixed paraffin embedded material. *PLoS One*. 2016; 11:e0149405. [PubMed: 26919633]
39. Mafficini A, Amato E, Fassan M, Simbolo M, Antonello D, Vicentini C, et al. Reporting tumor molecular heterogeneity in histopathological diagnosis. *PLoS One*. 2014; 9:e104979. [PubMed: 25127237]
40. Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. *Nature*. 2013; 501:355–364. [PubMed: 24048068]

HIGHLIGHTS

- We designed a prediction model to stratify endometrial cancer patients by risk levels using somatic mutations from TCGA.
- The prediction model including variant allele frequencies for each somatic gene mutation was superior to any other strategy.
- Stratifying patients accordingly to risk could individualize cancer treatment before and after surgery.

Fig. 1.

Flow diagram delineating the strategies and methods used for the analysis. Flow diagram detailing number of patients included in each group, methods and software used, and variable selection approaches in this study. Lasso: least absolute shrinkage and selection operator. CMA and glmnet: prediction and classification software packages.

Fig. 2.

AUC for prediction of levels of risk in endometrial cancer in the strategy #2 and comparison with clinical prediction models. A. Box plot representation of AUCs for different methods used in the 411-gene *complete prediction model* and measured in terms of AUC (y axis): RF: Random Forest; LASSO: least absolute shrinkage and selection operator; ELASNET: Elastic Net; DLDA: Diagonal Discriminant Analysis. PLS-LR: PLS - logistic regression; comBOOST: Component-wise Boosting; GBM: Tree-based Boosting; PLR: Penalized Logistic Regression; PLS: Partial Least Squares; PLS-RF: Positive Random Forest. B. Box plot representation of AUCs (y axis) for different methods used in *optimized prediction model* with selected 35 genes (same methods) C. Box plot representation of AUCs (y axis) for clinical prediction models including age and tumor grade (same methods).

Fig. 3. Mean AUC analysis using variant allele frequency for each somatic mutation. Mean AUC based on variant allele frequency of 53 most informative mutations.

Fig. 4. Genomic position and coefficients of 53 genes selected in the prediction model using variant allele frequency. A. 53 most informative mutated genes of the best prediction model with: Chromosomal arm location: where the gene is located, p is short and q is long; Allele frequencies range: range of the mutation allele frequencies; Coefficients: weight or importance of this gene in the prediction model. Human genome version hg19. B. From external to internal: Chromosome bands: circular representation of all chromosomes (centromere is in red); somatic gene mutations (violet) placed at their chromosomal location; gene expression: differential expression between high risk and low risk (HR/LR) for those mutated genes in TCGA database (red is over-expressed, green is under-expressed).

Table 1

Comparison of AUCs and CIs of the clinical prediction model with prediction models including mutations. Clinical prediction models included age and tumor grade, the only two clinical variables that are associated with level of risk and are available before surgery. The clinical prediction model was performed with the same software and methods as the prediction models using mutations.

	AUC	SE	95% CI
Random Forest	0.82	0.02	0.78, 0.87
Lasso	0.75	0.03	0.70, 0.80
Elastic Net	0.75	0.03	0.70, 0.80
Diagonal Discriminant Analysis	0.82	0.02	0.77, 0.86
PLS - logistic regression	0.82	0.02	0.78, 0.87
Component-wise Boosting	0.70	0.03	0.64, 0.77
Tree-based Boosting	0.84	0.02	0.81, 0.88
Penalized Logistic Regression	0.82	0.02	0.78, 0.87
Partial Least Squares	0.82	0.02	0.78, 0.87
PLS - Random Forest	0.78	0.02	0.75, 0.81

AUC: area under the ROC curve. CI: confidence intervals. SE: Standard Error. PLS: Partial least squares; Lasso: least absolute shrinkage and selection operator.