ARTIFICIAL INTELLIGENCE

original reports abstract

Analytical Validation of a Deep Neural Network Algorithm for the Detection of Ovarian Cancer

Gerard Reilly, MD[1]; Rowan G. Bullock, BS[2]; Jessica Greenwood, MS, CGC[2]; Daniel R. Ure, MS[2]; Erin Stewart, MS[2]; Pierre Davidoff, MS[2]; Justin DeGrazia, BS[2]; Herbert Fritsche, PhD[2]; Charles J. Dunton, MD[2]; Nitin Bhardwaj, PhD[2]; and Lesley E. Northrop, PhD[2]

PURPOSE Early detection of ovarian cancer, the deadliest gynecologic cancer, is crucial for reducing mortality. Current noninvasive risk assessment measures include protein biomarkers in combination with other clinical factors, which vary in their accuracy. Machine learning can be applied to optimizing the combination of these features, leading to more accurate assessment of malignancy. However, the low prevalence of the disease can make rigorous validation of these tests challenging and can result in unbalanced performance.

METHODS MIA3G is a deep feedforward neural network for ovarian cancer risk assessment, using seven protein biomarkers along with age and menopausal status as input features. The algorithm was developed on a heterogenous data set of 1,067 serum specimens from women with adnexal masses (prevalence = 31.8%). It was subsequently validated on a cohort almost twice that size (N = 2,000).

RESULTS In the analytical validation data set (prevalence = 4.9%), MIA3G demonstrated a sensitivity of 89.8% and a specificity of 84.02%. The positive predictive value was 22.45%, and the negative predictive value was 99.38%. When stratified by cancer type and stage, MIA3G achieved sensitivities of 94.94% for epithelial ovarian cancer, 76.92% for early-stage cancer, and 98.04% for late-stage cancer.

CONCLUSION The balanced performance of MIA3G leads to a high sensitivity and high specificity, a combination that may be clinically useful for providers in evaluating the appropriate management strategy for their patients. Limitations of this work include the largely retrospective nature of the data set and the unequal, albeit random, assignment of histologic subtypes between the training and validation data sets. Future directions may include the addition of new biomarkers or other modalities to strengthen the performance of the algorithm.

JCO Clin Cancer Inform 6:e2100192. © 2022 by American Society of Clinical Oncology

INTRODUCTION

Adnexal masses are a common gynecologic condition. With approximately 10% of women undergoing surgery for an adnexal mass during their lifetime, the research efforts to date have focused on tools designed to identify which of these masses are cancerous.[1,2] Ovarian cancer is the deadliest gynecologic cancer, and therefore, prompt and correct identification of malignancies is crucial. However, the incidence of ovarian cancer is still relatively low.[3] Approximately 85% of masses in premenopausal women will be benign, so testing that can accurately differentiate malignant masses from those that require less extensive intervention and treatment is of clinical value.[1]

Identification of a pelvic mass may occur during physical examination but more likely via imaging, typically with transvaginal ultrasonography. Biopsy is usually avoided to reduce the risk of disrupting the cyst wall and allowing any potential malignant cells to disseminate.[4] When a mass shows clear indications of malignancy, the patient benefits from appropriate referral to a gynecologic oncologist for surgery, staging, and any further treatment.[5]

Beyond imaging, additional methods of assessing adnexal masses include the use of biomarker-based blood tests, such as cancer antigen 125 (CA125) and human epididymis protein 4 (HE4). Relying on these traditional methods to stratify the oncologic risk of adnexal masses has several challenges. First, a small set of biomarkers may not be able to ascertain the physiology of certain ovarian cancers because different histologic subtypes are known to present with different biomarker patterns.[6-8] Second, the process of using a set threshold for each biomarker can become cumbersome when multiple markers are added to the analysis. Third, this process may be further complicated by the age and menopausal status of the patient, which can affect the baseline or so-called normal level of these proteins.

Machine learning–based classification models can address these limitations, which is why their use in

ASSOCIATED CONTENT

Appendix

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on April 20, 2022 and published at ascopubs.org/journal/cci on June 7, 2022: DOI https://doi.org/10.1200/CCI.21.00192

CONTEXT

Key Objective

Our objective was to examine the potential of a noninvasive machine learning tool to accurately assess the risk of ovarian malignancy in patients with pelvic masses.

Knowledge Generated

The deep neural network was trained on a large heterogenous data set obtained from patients who had presented with adnexal masses and used seven serum proteins, age, and menopausal status as inputs. In the analytical validity data set, which simulated real-world prevalence for ovarian malignancy (4.9%), the algorithm demonstrated a sensitivity of 89.8%, a specificity of 84.0%, a positive predictive value of 22.5%, and a negative predictive value of 99.5%.

Relevance

Ovarian cancer is the deadliest gynecologic cancer, and most cases are diagnosed at a late stage, which has low survival rates. Current noninvasive risk assessment measures vary in their accuracy, so the balanced sensitivity and specificity of this algorithm will be a clinically useful combination for providers evaluating appropriate care strategies for patients presenting with a pelvic mass.

early cancer detection and risk stratification is increasing.[9] These models are capable of incorporating a long list of protein biomarkers along with clinical/health features as inputs to generate a unified score for risk assessment. However, building these models can be challenging because of the low incidence of ovarian cancer. Having a small set of positive samples for training can result in a skewed model with a high specificity but a low sensitivity. Developing a balanced classification model with high sensitivity and specificity is crucial, especially given the mortality implications of false negatives (FNs) and the burden on the health care system and the patient of false positives (FPs).

This study describes the development and validation process used to establish test performance metrics for MIA3G, a new machine learning algorithm to assess ovarian cancer risk in patients with an adnexal mass. Powered by a robust data set inclusive of a large number of malignancies for training and testing, this algorithm has demonstrated balanced performance in a large analytical validation set.

METHODS

Algorithm Description

The MIA3G assay is an algorithm developed with a proprietary application of machine learning methods whose purpose is to stratify women with an ovarian mass into two categories—low and elevated risk of malignancy. The algorithm uses supervised learning with known histopathology diagnoses (malignant and nonmalignant) as the labels for algorithm training. MIA3G is a classification deep feedforward neural network that uses the following features as inputs: age, menopausal status, and seven protein biomarker values for each patient. The neural network has multiple hidden layers each with their own weighted nodes and activation functions. The neural network is regularized using node dropout to reduce overfitting where a percentage of the nodes are randomly omitted from each hidden layer during training.[10] The final layer of the neural network has two nodes and uses the *softmax* function to assign a binary classification: low or elevated risk of malignancy. Additional details of methods used to reduce overfitting and oversampling are provided section S1.1. of Appendix 1.

Protein Biomarkers and Input Features

Seven biomarkers are used in the MIA3G algorithm: CA125, HE4, beta-2 microglobulin, apolipoprotein A-1, transferrin, transthyretin, and follicle-stimulating hormone. CA125 and HE4 were chosen for their overexpression in many types of ovarian cancers.[11,12] The remaining biomarkers have demonstrated ability to detect malignancy in patients with low serum CA125 and/or HE4, such as early-stage malignancies, as well as reducing FPs in benign cases for which serum CA125 and/or HE4 were elevated for other reasons.[13-16] These features have been examined for their correlation with each other and their contribution (Appendix Fig A1). Biomarker assays are performed using the Roche cobas 6000 analyzer, according to the manufacturer's instructions for use (*Roche Corporation*, Pleasanton, CA). In addition to these biomarkers, the patient's age and menopausal status are used as categorical input features. Menopause is defined as the absence of menses for $\geq$ 12 months.

Studies and Sample Sets

To create a highly generalizable classification algorithm, it is essential to train it on a diverse set of specimens with a wide reference range of biomarkers and other features. To this end, a heterogenous set of specimens was first created by combining samples from several prospective and retrospective studies, all of which underwent Institutional Review Board approval and in accordance with appropriate regulatory and ethical guidelines (Table 1).

TABLE 1. Sample Set Composition

Study	IRB/Protocol No.	No.
OVAWatch Prospective Clinical Study[17]	RP 08-2020, RP 05-2019, RP 04-2019	35
Aspira Specimen (Serum) Bank	RP 01-2016/Pro00027159	290
OVA1 Postmarket Study[18]	OVA1-PS1-CO4	1,385
OVA500 Study[19]	OVA2-002-CO3	511
University of Washington Study[20]	OVA1-7788	218
OVA1 Study[21]	OVA1-001-CO1	574
BioBank[22]	SHARE v5.2 10.May.2021, IRB#: 2017-198	54
Total		3,067

Abbreviation: IRB, Institutional Review Board.

Broadly, the inclusion criteria for these studies were as follows:

- Patient age ≥ 18 years
- Informed consent provided by the patient to participate in research
- Patient agreeable to phlebotomy
- Patient had a documented pelvic mass that was planned for surgical intervention within 3 months of imaging. The pelvic mass was confirmed by imaging (computed tomography, ultrasonography, or magnetic resonance imaging) before enrollment.

Exclusion criteria included a diagnosis of malignancy in the previous 5 years (except nonmelanoma skin cancers). Exclusion criteria also included pelvic surgery within six weeks before enrollment in the study.

This heterogenous set comprised a total of 3,067 samples (Fig 1). The composite set was randomly broken into two nonoverlapping sets such that

- One thousand sixty-seven samples were used for development of the algorithm and formed the training and testing set.
- The remaining 2,000 samples were used for analytical validation.
- Each set roughly received samples from every study proportionate to the size of the study.
- The validation set had a prevalence rate of approximately 5% (98 malignant and 1,902 benign samples).

Although the sample size and prevalence of malignancy were fixed, the sample assignment to each set was completely random, performed using a random number generator to remove any potential bias. The above binning of samples into development and validation sets ensures that not only is the assignment of samples fair and random, but it also allows the algorithm to be trained/tested and then validated on sets that have an optimal level of similarities (and differences). Table 2 details the clinicopathologic makeup of each set including age, pathology, histologic subtypes, and stages.

Data and Ethics

All data were obtained from Institutional Review Board–approved trials, from adult patients who provided informed consent to participate in research. Data obtained in this analysis are proprietary to Aspira Women's Health Inc.

Training and Testing

The MIA3G algorithm was developed on 1,067 specimens composed of proportionate samples from every study

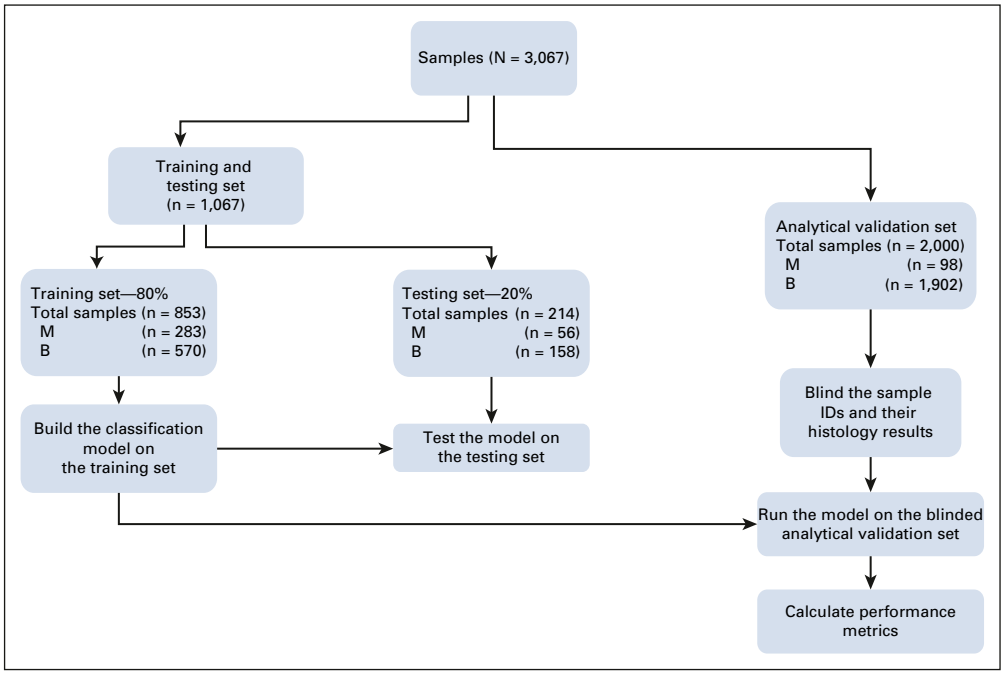

FIG 1. Workflow of the development and validation of the algorithm. B, benign samples; M, malignant samples.

TABLE 2. Clinicopathologic Breakdown of Training, Test, and Validation Data Sets

	Training Set			Test Set			Validation Set		
Menopausal Status	**All (n = 853)**	**Pre (n = 410)**	**Post (n = 443)**	**All (n = 214)**	**Pre (n = 105)**	**Post (n = 109)**	**All (n = 2,000)**	**Pre (n = 1,193)**	**Post (n = 807)**
Age, years, mean	51.3	40.3	61.4	50.8	40.5	60.8	47.5	39.5	59.4
Pathology diagnosis, No. (%)									
Benign ovarian conditions	548 (64.2)	319 (77.8)	229 (51.7)	152 (71.0)	86 (81.9)	66 (60.6)	1,836 (91.8)	1,136 (95.2)	700 (86.7)
Low malignant potential (borderline)	25 (2.9)	9 (2.2)	16 (3.6)	6 (2.8)	1 (1.0)	5 (4.6)	66 (3.3)	31 (2.6)	35 (4.3)
Epithelial primary ovarian cancer	200 (23.4)	49 (12.0)	151 (34.1)	45 (21.0)	12 (11.4)	33 (30.3)	79 (4.0)	18 (1.5)	61 (7.6)
Nonepithelial primary ovarian cancer	41 (4.8)	19 (4.6)	22 (5.0)	5 (2.3)	3 (2.9)	2 (1.8)	6 (0.3)	4 (0.3)	2 (0.2)
Non-primary malignancies	39 (4.6)	14 (3.4)	25 (5.6)	6 (2.8)	3 (2.9)	3 (2.8)	13 (0.7	4 (0.3)	9 (1.1)
Stage (primary ovarian malignancies), No. (%)									
Stage I	90 (37.3)	30 (44.1)	60 (34.7)	15 (30.0)	6 (40.0)	9 (25.7)	16 (18.8)	7 (31.8)	9 (14.3)
Stage II	33 (13.7)	10 (14.7)	23 (13.3)	5 (10.0)	1 (6.7)	4 (11.4)	10 (11.8)	4 (18.2)	6 (9.5)
Stage III	83 (34.4)	16 (23.5)	67 (38.7)	24 (48.0)	7 (46.7)	17 (48.6)	46 (54.1)	8 (36.4)	38 (60.3)
Stage IV	17 (7.1)	4 (5.9)	13 (7.5)	4 (8.0)	0 (0.0)	4 (11.4)	5 (5.9)	0 (0.0)	5 (7.9)
Not staged	18 (7.5)	8 (11.8)	10 (5.8)	2 (4.0)	1 (6.7)	1 (2.9)	8 (9.4)	3 (13.6)	5 (7.9)
Histologic subtype (primary ovarian malignancies), No. (%)									
EOC									
Serous	105 (43.6)	21 (30.9)	84 (48.6)	25 (50.0)	6 (40.0)	19 (54.3)	46 (54.1)	7 (31.8)	39 (61.9)
Endometrioid	31 (12.9)	10 (14.7)	21 (12.1)	5 (10.0)	4 (26.7)	1 (2.9)	10 (11.8)	3 (13.6)	7 (11.1)
Mucinous	21 (8.7)	9 (13.2)	12 (6.9)	7 (14.0)	1 (6.7)	6 (17.1)	6 (7.1)	2 (9.1)	4 (6.3)
Clear cell	18 (7.5)	4 (5.9)	14 (8.1)	3 (6.0)	1 (6.7)	2 (5.7)	11 (12.9)	4 (18.2)	7 (11.1)
Mixed	12 (5.0)	2 (2.9)	10 (5.8)	4 (8.0)	0 (0.0)	4 (11.4)	2 (2.4)	1 (4.5)	1 (1.6)
Poorly differentiated	6 (2.5)	1 (1.5)	5 (2.9)	1 (2.0)	0 (0.0)	1 (2.9)	3 (3.5)	1 (4.5)	2 (3.2)
Transitional cell	3 (1.2)	1 (1.5)	2 (1.2)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Others	4 (1.7)	1 (1.5)	3 (1.7)	0 (0.0)	0 (0.0)	0 (0.0)	1 (1.2)	0 (0.0)	1 (1.6)
Non-EOC									
Sex cord stromal	20 (8.3)	10 (14.7)	10 (5.8)	1 (2.0)	1 (6.7)	0 (0.0)	5 (5.9)	3 (13.6)	2 (3.2)
Germ cell	11 (4.6)	8 (11.8)	3 (1.7)	0 (0.0)	0 (0.0)	0 (0.0)	1 (1.2)	1 (4.5)	0 (0.0)
Sarcoma/carcinosarcoma	9 (3.7)	0 (0.0)	9 (5.2)	4 (8.0)	2 (13.3)	2 (5.7)	0 (0.0)	0 (0.0)	0 (0.0)
Others	1 (0.4)	1 (1.5)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

Abbreviation: EOC, epithelial ovarian cancer.

(Fig 1), with 339 malignant and 728 benign samples resulting in a prevalence of 31%. This set was randomly divided into a training set (n = 853) and a nonoverlapping testing set (n = 214), representing 80% and 20% of the available samples, respectively. The algorithm was built on the training set and tested on the testing set to obtain an initial assessment of its performance. The performance metrics for the test data set are provided in Appendix Table A1.

The numbers of malignant and benign specimens were further balanced for algorithm training using an adaptation of the synthetic minority oversampling technique (SMOTE) that balances the minority and majority classes by creating synthetic observations near the decision boundary (called Borderline-SMOTE).[23] The resulting data set has an equivalent number of malignant and benign specimens, where the synthetic observations are close to the decision boundary. In the case of MIA3G, the synthetic observations improve the algorithm's ability to discern between malignant and benign specimens. To improve malignancy detection, a modestly higher weight was attached to the positive class during algorithm training in MIA3G. Weighing the malignant samples during training improved on the gains from balancing using the Borderline-SMOTE in positive detection, while having a negligible impact on benign discernment.

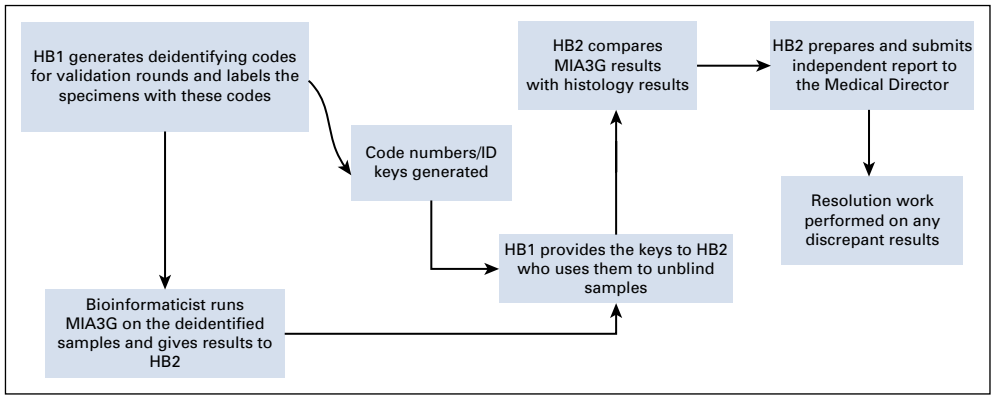

FIG 2. Workflow of the analytical validation exercise. HB, honest broker.

Several algorithms and software libraries were used to explore which technique would return the best risk classification for ovarian cancer. The caret library in R was used to screen 190 classification algorithms on the data.[24] Most algorithms in the caret library did not successfully classify ovarian cancer with a high level of sensitivity. Deep feed-forward neural networks demonstrated a high and balanced sensitivity, negative predictive value (NPV), and specificity, leading to the selection of this algorithm for the development of MIA3G. Network hyperparameters evaluated during algorithm training and testing included the following: network architectures, activation functions, loss functions, node dropout for algorithm regularization, and learning rates. The final MIA3G algorithm is a network with these hyperparameters optimized to stratify malignancy risk. This algorithm was locked and used for subsequent analytical validation.

Analytical Performance Validation

Analytical validation was performed on 2,000 samples with 98 malignant and 1,902 benign specimens, resulting in a prevalence of 4.9%. Once the algorithm was developed and locked in a cloud-based Health Insurance Portability and Accountability Act–compliant infrastructure, it was then run on the analytical validation samples in a blinded manner so that the person running the algorithm was blinded to the sample identities and their pathology results. Two honest brokers (HB1 and HB2) were used to deidentify the samples, run the algorithm blinded, compare the classification of samples with the histology results, and then issue

TABLE 3. Performance of MIA3G in the Validation Data Set

Group	Malignant	Benign	TP	TN	FP	FN	Sens (%)	Spec (%)	PPV (%)	NPV (%)
All	98	1,902	88	1,598	304	10	89.80	84.02	22.45	99.38
Premenopausal	26	1,167	21	1,072	95	5	80.77	91.86	18.10	99.54
Postmenopausal	72	735	67	526	209	5	93.06	71.56	24.28	99.06
EOC	79	—	75	—	—	4	94.94	—	—	—
Non-EOC	6	—	1	—	—	5	16.67	—	—	—
Stage I	16	—	11	—	—	5	68.75	—	—	—
Stage II	10	—	9	—	—	1	90.00	—	—	—
Stage III	46	—	45	—	—	1	97.83	—	—	—
Stage IV	5	—	5	—	—	0	100.00	—	—	—
Early stage (I and II)	26	—	20	—	—	6	76.92	—	—	—
Late stage (III and IV)	51	—	50	—	—	1	98.04	—	—	—
Not staged	8	—	6	—	—	2	75.00	—	—	—
Nonprimary	13	—	12	—	—	1	92.31	—	—	—
LMP	—	66	—	33	33	—	—	50.00	—	—
Other benigns	—	1,836	—	1,565	271	—	—	85.24	—	—

NOTE. The number of cases or metrics not applicable for that category are displayed by —.

Abbreviations: EOC, epithelial ovarian cancer; FN, false negative; FP, false positive; LMP, low malignant potential/borderline tumor; NPV, negative predictive value; PPV, positive predictive value; Sens, sensitivity; Spec, specificity; TN, true negative; TP, true positive.

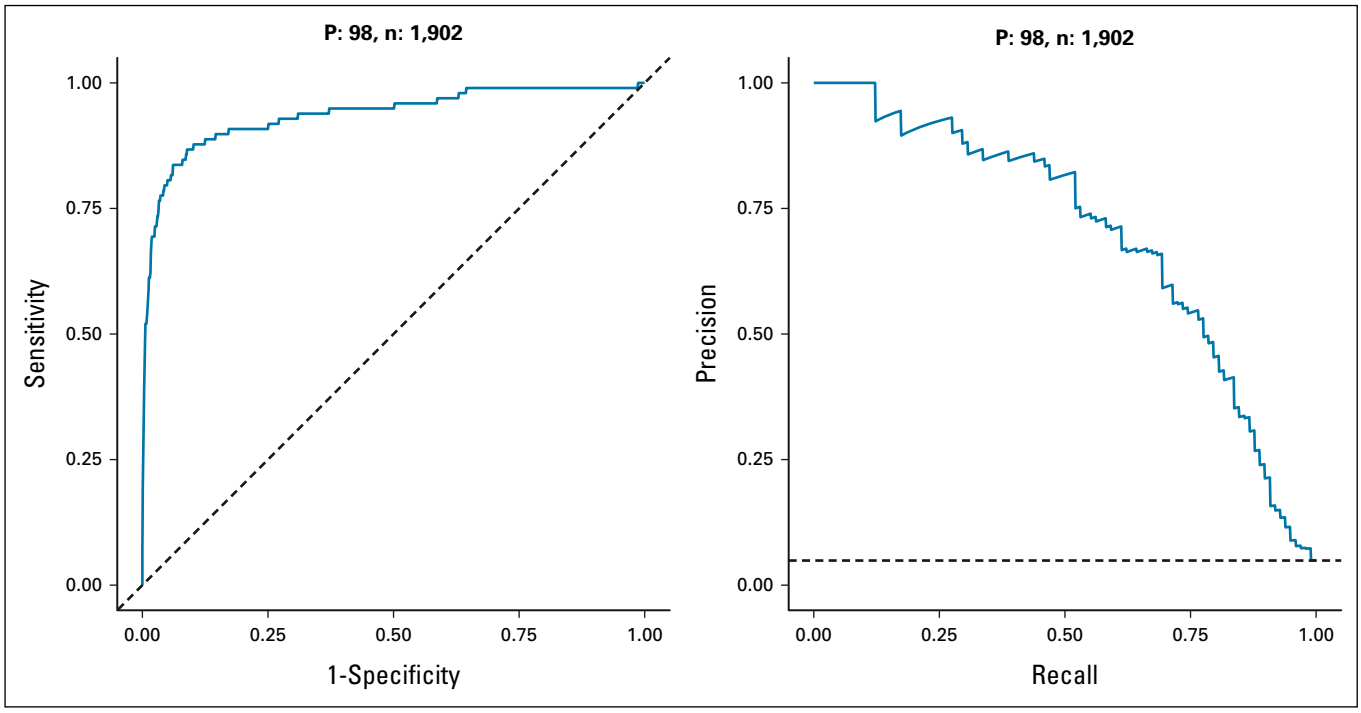

FIG 3. ROC and precision-recall curves for the algorithm. Area under the receiver operating characteristic curve: 0.938, area under the precision-recall curve: 0.700. n, negative. P, positive; ROC, receiver operating characteristic.

an independent report containing performance metrics on the basis of their findings (Fig 2).

RESULTS

Performance metrics along with counts of true positives, true negatives, FPs, and FNs from analytical validation are provided in Table 3. Receiver operating characteristic and precision-recall curves are also plotted (Fig 3). Overall, a sensitivity of 89.8% and a specificity of 84.02% were achieved, with an area under the curve value of 0.938. MIA3G demonstrated an NPV of 99.38%. The positive predictive value was lower at 22.45% because of the low prevalence of disease (approximately 5%) in this data set. Metrics have also been provided for specimens stratified by menopausal status, cancer stage, cancer type, and malignancy potential. MIA3G was able to detect 20 of 26 early-stage cancers (76.92% sensitivity) and misclassified only one late-stage malignancy (98.04% sensitivity). The algorithm also correctly classified nine of the 10 metastatic ovarian cancer cases (90% sensitivity) and 75 of 79 instances of epithelial ovarian cancer, the most common type of ovarian cancer (94.94% sensitivity).

DISCUSSION

A thorough and rigorous development process combined with comprehensive analytical validation is the cornerstone of any clinical laboratory–developed test. It is the foundation for setting quality standards and illustrates the performance and reliability of the underlying machinery. MIA3G has undergone a rigorous and blinded analytical validation process that meets the highest regulatory standards in evaluating all aspects of the test.

After assessing several classification algorithms, MIA3G was trained on neural networks with the most balanced performance and then tested on a heterogeneous cohort. The model was optimized to reduce overfitting, and an oversampling technique was used to achieve a balanced performance, which was higher than all other methods that were explored (Appendix Table A2). The training and testing stage used $>$ 1,050 specimens with $>$ 30% positive specimens indicative of a high-risk ovarian cancer population. This development was followed by a detailed validation process on 2,000 specimens that show performance in a low prevalence population (approximately 5%), making the algorithm highly generalizable. MIA3G has also been meticulously validated for its repeatability and reproducibility (Appendix Table A3).

The potential clinical utility of MIA3G in the evaluation of adnexal masses comes from its balanced performance, which is facilitated by three development features: a large malignant set used in training and testing (n = 339), the SMOTE technique applied to further boost the positive set, and a higher weight attached to the positive class. These features lead to an algorithm with a high sensitivity, a vital feature that shows the high mortality of ovarian cancer, while retaining a high specificity. The high specificity drives a high NPV in a population with a lower disease prevalence where clinical management options may include conservative management and at the same time minimizes the

potentially lethal implications of FNs in the context of cancer detection.

Limitations of this study include the nature of the development data set. Although MIA3G was developed and validated on a highly diverse cohort obtained by merging several studies, most of these studies were retrospective in design with data collected from patients who were confirmed to have an adnexal mass and scheduled for surgery at the time of diagnosis. To address this, prospective trials are currently underway to validate the algorithm's performance in patients with a variety of clinical presentations.

In addition, because of the random assignment of samples to the training and validation data sets, there was no way to match the distribution of cancer types between sets (Table 2). For example, by happenstance, five of the tumors in the validation set were stromal tumors and one was a germ cell tumor, subtypes known to have a different biomarker presentation compared with the more common epithelial types. In the test set, however, MIA3G demonstrated 100% sensitivity in nonepithelial malignancies, as sarcomas and carcinosarcomas comprised 4 of 5 nonepithelial malignancies in that set (Appendix Table A1). These cancer types present more similar to epithelial ovarian cancer in terms of biomarker distribution. Nonepithelial subtypes are rare presentations of ovarian cancer, comprising approximately 10% of all ovarian malignancies,[25] so their particularly low incidence presents a challenge with regard to generating sufficient data for training and validating machine learning algorithms. Future directions include evaluating how to train an algorithm on multiple subtypes that express different biomarker patterns and achieve consistent test performance across these subtypes.

The application of a deep neural network algorithm to biomarker testing opens significant areas for future study. Understanding where the algorithm fails provides an opportunity for deeper exploration into alternate biologic explanations for FP and FN results. For example, there is a possibility that some combination of biomarkers may be identifying cancers outside of the ovaries and therefore correctly suggesting malignancy, albeit not of ovarian origin. As a step for improvement, expanding the number and types of features that feed into the algorithm may help further enhance the sensitivity and specificity of the test. Preliminary efforts are underway to evaluate the addition of novel biomarkers and other modalities such as microRNA, cell tumor DNA, and other genomic identifiers that may strengthen the algorithms' ability to both detect and rule out malignancy and advance the diagnostic ability of noninvasive testing.

AFFILIATIONS

[1]Axia Women's Health, Cincinnati, OH
[2]Aspira Women's Health, Trumbull, CT

CORRESPONDING AUTHOR

Rowan G. Bullock, BS, Aspira Women's Health, Inc, 12117 Bee Caves Rd #3-100, Austin, TX 78738; e-mail: rbullock@aspirawh.com.

PRIOR PRESENTATION

Presented at the 2021 ASCO meeting (abstr 339987), June 4-8, 2021, virtual.

DATA SHARING STATEMENT

Individual participant data from clinical trials privately funded by Aspira Women's Health Inc (AWH) are proprietary to AWH and will not be made available. IRB/protocol numbers for these include RP 08-2020, RP 05-2019, RP 04-2019, RP 01-2016, OVA1-PS1-CO4, OVA2-002-CO3, OVA1-7788, and OVA1-001-CO1. Individual participant data from the SHARE Biorepository are managed by Spectrum Health.

AUTHOR CONTRIBUTIONS

Conception and design: Gerard Reilly, Rowan G. Bullock, Jessica Greenwood, Daniel R. Ure, Justin DeGrazia, Herbert Fritsche, Nitin Bhardwaj, Lesley E. Northrop
Administrative support: Herbert Fritsche
Provision of study materials or patients: Herbert Fritsche
Collection and assembly of data: Rowan G. Bullock, Daniel R. Ure, Erin Stewart, Pierre Davidoff, Justin DeGrazia
Data analysis and interpretation: Rowan G. Bullock, Jessica Greenwood, Daniel R. Ure, Pierre Davidoff, Herbert Fritsche, Charles J. Dunton, Nitin Bhardwaj, Lesley E. Northrop
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Gerard Reilly
Employment: Axia Women's Health
Leadership: Axia women's Health
Stock and Other Ownership Interests: Axia Women's Health
Consulting or Advisory Role: Meditrina, Early Bird, ChannelMed
Research Funding: AbbVie, Aspira Labs, Ravel, Natera, Bayer, Prima-Temp, MicroCube, Evofem

Rowan G. Bullock
Employment: Aspira Women's Health

Jessica Greenwood
Employment: Aspira Women's Health
Stock and Other Ownership Interests: Aspira Women's Health
Consulting or Advisory Role: Delphi Diagnostics, MyHeritage
Travel, Accommodations, Expenses: Aspira Women's Health

Daniel R. Ure
Employment: Aspira Women's Health

Stock and Other Ownership Interests: Aspira Women's Health
Travel, Accommodations, Expenses: Aspira Women's Health

Erin Stewart
Employment: Aspira Women's Health, SpeeDX Inc
Stock and Other Ownership Interests: Aspira Women's Health
Consulting or Advisory Role: Aspira Women's Health

Pierre Davidoff
Employment: Celmatix
Stock and Other Ownership Interests: Aspira Women's Health
Consulting or Advisory Role: Aspira Women's Health

Justin DeGrazia
Employment: Aspira Women's Health

Herbert Fritsche
Employment: Aspira Lab, Brevitest Lab
Stock and Other Ownership Interests: Oncimmune Ltd
Consulting or Advisory Role: EDP
Patents, Royalties, Other Intellectual Property: Hold several patents, none of which are owned by me. Rights assigned to the University of Texas, Aspira Labs or EDP
Travel, Accommodations, Expenses: Aspira Lab and EDP

Charles J. Dunton
Employment: Aspira Women's Health
Leadership: Aspira Women's Health
Stock and Other Ownership Interests: Aspira Women's Health
Consulting or Advisory Role: Aspira Women's Health

Nitin Bhardwaj
Employment: Aspira Women's Health
Stock and Other Ownership Interests: Aspira Women's Health
Consulting or Advisory Role: Aspira Women's Health

Lesley E. Northrop
Employment: Aspira Women's Health
Leadership: Aspira Women's Health
Stock and Other Ownership Interests: Aspira Women's Health

No other potential conflicts of interest were reported.

REFERENCES

1. Ueland FR, Fredericks TI: Ovarian masses: Surgery or surveillance? OBG Manag 30:17-24, 26, 2018
2. Rim SH, Hirsch S, Thomas CC, et al: Gynecologic oncologists involvement on ovarian cancer standard of care receipt and survival. World J Obstet Gynecol 5:187-196, 2016
3. Cancer of the ovary—Cancer stat facts. SEER: Surveillance, Epidemiology, and End Results program. 2021. https://seer.cancer.gov/statfacts/html/ovary.html
4. May T, Oza A: Conservative management of adnexal masses. Lancet Oncol 20:326-327, 2019
5. Doubeni CA, Doubeni AR, Myers AE: Diagnosis and management of ovarian cancer. Am Fam Physician 93:937-944, 2016
6. Badgwell D, Bast RC Jr: Early detection of ovarian cancer. Dis Markers 23:397-410, 2007
7. Choi JH, Sohn GS, Chay DB, et al: Preoperative serum levels of cancer antigen 125 and carcinoembryonic antigen ratio can improve differentiation between mucinous ovarian carcinoma and other epithelial ovarian carcinomas. Obstet Gynecol Sci 61:344-351, 2018
8. Kommoss F, Lehr HA: Sex cord-stromal tumors of the ovary: Current aspects with a focus on granulosa cell tumors, Sertoli-Leydig cell tumors, and gynandroblastomas [in German]. Pathologe 40:61-72, 2019
9. Artificial Intelligence. National Cancer Institute, 2020. https://www.cancer.gov/research/areas/diagnosis/artificial-intelligence
10. Srivastava N, Hinton G, Krizhevsky A, et al: Dropout: A simple way to prevent neural networks from overfitting. J Machine Learn Res 15:1929-1958, 2014
11. Bast RC Jr, Klug TL, St John E, et al: A radioimmunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer. N Engl J Med 309:883-887, 1983
12. Drapkin R, von Horsten HH, Lin Y, et al: Human epididymis protein 4 (HE4) is a secreted glycoprotein that is overexpressed by serous and endometrioid ovarian carcinomas. Cancer Res 65:2162-2169, 2005
13. Zhang Z, Bast RC Jr, Yu Y, et al: Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. Cancer Res 64:5882-5890, 2004
14. Yang HS, Li Y, Deng HX, et al: Identification of beta2-microglobulin as a potential target for ovarian cancer. Cancer Biol Ther 8:2323-2328, 2009
15. Kozak KR, Su F, Whitelegge JP, et al: Characterization of serum biomarkers for detection of early stage ovarian cancer. Proteomics 5:4589-4596, 2005
16. Zhang Z: An in vitro diagnostic multivariate index assay (IVDMIA) for ovarian cancer: Harvesting the power of multiple biomarkers. Rev Obstet Gynecol 5:35-41, 2012
17. Clinicaltrials.gov: A multivariate index assay for ovarian cancer risk assessment in women with adnexal mass and high-risk germline variants. 2021. https://clinicaltrials.gov/ct2/show/NCT04487405
18. Zhang Z, Bullock RG, Fritsche H: Adnexal mass risk assessment: A multivariate index assay for malignancy risk stratification. Future Oncol 15:3783-3795, 2019
19. Bristow RE, Smith A, Zhang Z, et al: Ovarian malignancy risk stratification of the adnexal mass using a multivariate index assay. Gynecol Oncol 128:252-259, 2013
20. Urban RR, Smith A, Agnew K, et al: Evaluation of a validated biomarker test in combination with a symptom index to predict ovarian malignancy. Int J Gynecol Cancer 27:233-238, 2017
21. Clinicaltrials.gov: Whole blood collection protocol for ovarian assay clinical trial in women with ovarian tumors. 2008. https://clinicaltrials.gov/ct2/show/NCT00436189
22. Open Clinical Trials: Other. Spectrum Health, 2019. https://www.spectrumhealth.org/-/media/spectrumhealth/documents/clinical-research/available-clinical-trials-pdfs/other-health-clinical-trials/other-open-clinical-trials.pdf
23. Han H, Wang WY, Mao BH, et al: Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning, in Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, Volume 3644. Berlin, Heidelberg, Springer, 2005
24. Kuhn M, Wing J, Weston S, et al: Package "Caret". 2021. https://cran.r-project.org/web/packages/caret/caret.pdf
25. Berek JS, Bast RC Jr: Nonepithelial ovarian cancer, in Kufe DW, Pollock RE, Weichselbaum RR, et al (eds): Holland-Frei Cancer Medicine (ed 6). Hamilton, ON, Canada, BC Decker, 2003

APPENDIX

APPENDIX 1.

Supplementary 1: Methods

S1.1. Limiting overfitting and oversampling. Overfitting during model building was mitigated using randomized node dropout. Node dropout randomly drops units with their connections from the neural network during training. This prevents units from coadapting too much, and excess weight is given to specific nodes. This significantly reduces overfitting and gives major improvements over other regularization methods (Hinton G, et al: J Machine Learn Res 15:1929-1958, 2014).

We also adopted the BLS synthetic minority oversampling technique (SMOTE) for this purpose. Experiments conducted by Han et al have shown that the BLS-SMOTE approach achieves a better true-positive rate and F-value than SMOTE and random oversampling methods when working with imbalanced data. For every minority example, its *k* nearest neighbors from the same class are identified, then some examples are randomly selected from them according to the oversampling rate (Han H, et al, in Huang DS, et al [eds]: Volume Part I. Berlin, Heidelberg, Springer-Verlag, 2005, pp 878-887). After that, new synthetic examples are generated along the line between the minority example and its selected nearest neighbors. Unlike the existing oversampling methods, BLS-SMOTE oversamples the borderline minority examples, which in many cases, in our cohort, are early-stage cancers.

S1.2. Feature selection

S1.2.1. Feature correlation. We examined all the features for any strong correlation in the context of our cohort by performing a correlation analysis between the features (Appendix Fig A1).

Quite understandably, age and menopausal status are highest correlated features that the algorithm uses followed by follicle-stimulating hormone (FSH) protein biomarker, which is correlated with age and menopausal status. Removing menopausal status led to a modest decrease in algorithm performance metrics (n = 10 different data random assessments, a mean decrease of 5.8% in sensitivity in the test data, specificity remained equivalent). Although age and menopausal status are correlated, it was deemed worth including both for the retention of sensitivity in algorithm performance shown in the test data set.

Similarly, removing FSH led to roughly equivalent sensitivity; however, there was a 3.5% decrease in specificity. Again, it was deemed worth including FSH for the retention of algorithm performance shown in the test data set. There were no other correlations in the data that were either ≥ 0.5 or ≤ -0.5.

S1.2.2. Feature contribution. We also assessed variable importance for each of the input features of the algorithm (Appendix Fig A2). Permutation-based variable importance analysis was used. As the permutations are stochastic, some variability can be anticipated in the resulting importance depending on data seeding. The plot below is the representation of the mean 25 data random assignment seeds. Human epididymis protein 4, cancer antigen 125, menopausal status, and apolipoprotein A-1 age were the four most important features. These data along with the information from the correlation exploration suggest that all biomarkers and input variables are contributing in a meaningful manner to a variable extent.

Supplementary 2: Results

S2.1. Other machine learning methods. We also evaluated many other algorithms for their performance on the same data set. Appendix Table A1 lists the performance of other machine learning algorithms included in this analysis. Neural networks show the highest sensitivity and negative predictive value (the two metrics that we optimized so as to minimize false negatives, a decision on the basis of the high mortality of ovarian cancer, particularly when discovered at a late stage).

Supplementary 3: Precision

The MIA3G algorithm and individual analyte concentration measurements were rigorously evaluated for precision, that is, repeatability and reproducibility according to Clinical and Laboratory Standards Institute standard EP05-A2 (Tholen DW, et al: Clinical and Laboratory Standards Institute, 2014, pp 1-39). The precision study for MIA3G was designed to establish its performance across and within runs, days, and operators. The exercise was configured to be run by two individual laboratory operators to assess the contribution of between-operator variability in MIA3G. Each sample was run in triplicate, at two separate times per day with a minimum of 2 hours apart to evaluate variability of MIA3G within and across runs (ie, intrareproducibility). In addition, this process was repeated across 4 days to evaluate within and across day deviations (ie, interreproducibility).

Repeatability and reproducibility of MIA3G probability risk score were quantified in terms of %CV (percentage of coefficient of variation). CV captures the extent of variability of data in relation to the mean of the population tested. It is the ratio of the standard deviation to the mean and is used for comparing the degree of variation from one data series with another, even if the means are drastically different from one another. A value of 10% CV or lower is a widely accepted degree of variability. Within experiment, %CV captures repeatability, and across experiment, %CV demonstrates reproducibility (aka precision). MIA3G %CV is provided in Appendix Table A2 for three metrics: runs, days, and operator. A low %CV (high repeatability and reproducibility) was demonstrated, with all values being below or around 10% CV. Individual biomarkers also confirmed low variability at all three levels measured (data not shown).

TABLE A1. Performance of MIA3G in the Test Data Set

Group	Malig	Benign	TP	TN	FP	FN	Sens (%)	Spec (%)	PPV (%)	NPV (%)
All	56	158	51	139	19	5	91.07	87.97	72.86	96.53
Premenopausal	18	87	16	83	4	2	88.89	95.40	80.00	97.65
Postmenopausal	38	71	35	56	15	3	92.11	78.87	70.00	94.92
EOC	45	—	42	—	—	3	93.33	—	—	—
Non-EOC	5	—	5	—	—	0	100.00	—	—	—
Stage I	15	—	12	—	—	3	80.00	—	—	—
Stage II	5	—	5	—	—	0	100.00	—	—	—
Stage III	24	—	24	—	—	0	100.00	—	—	—
Stage IV	4	—	4	—	—	0	100.00	—	—	—
Early stage (I and II)	20	—	17	—	—	3	85.00	—	—	—
Late stage (III and IV)	28	—	28	—	—	0	100.00	—	—	—
Not staged	2	—	2	—	—	0	100.00	—	—	—
Not primary to the ovary	6	—	4	—	—	2	66.67	—	—	—
LMP	—	6	—	3	3	—	—	50.00	—	—
Other benigns	—	152	—	136	16	—	—	89.47	—	—

NOTE. The number of cases or metrics not applicable for that category are displayed by —.

Abbreviations: EOC, epithelial ovarian cancer; FN, false negative; FP, false positive; LMP, low malignant potential/borderline tumor; NPV, negative predictive value; PPV, positive predictive value; Sens, sensitivity; Spec, specificity; TN, true negative; TP, true positive.

TABLE A2. Performance of Other Methods in Comparison With Neural Networks, Which Demonstrated Highest Sensitivity and NPV

Model	Sens	Spec	PPV	NPV
C5.0	82.65	91.06	32.27	99.03
Naive Bayesian classifier	72.45	88.49	24.48	98.42
Boosted logistic regression	86.73	81.13	19.14	99.16
SVM with linear kernel	83.67	82.54	19.81	98.99
Boosted smoothing spline	79.59	86.54	23.35	98.80
Generalized linear model	83.67	83.39	20.60	99.00
Self-organizing maps	77.17	80.54	16.10	98.65
Heteroscedastic discriminatory analysis	59.18	98.26	63.74	97.90
Neural network	89.80	84.02	22.45	99.38

Abbreviations: NPV, negative predictive value; PPV, positive predictive value; Sens, sensitivity; Spec, specificity; SVM, support vector machine.

TABLE A3. %CV Measurement of the MIA3G for Runs, Days, and Operators by Sample (pooled serum)

Metric	Sample		
Serum pool ID	25	26	27
MIA3G risk score			
%CV within runs	10.6	6.20	6.60
%CV across runs	0.0	0.00	0.90
%CV within days	10.7	6.30	6.80
%CV across days	0.0	3.00	3.60
%CV within operators	10.6	6.20	6.70
%CV across operators	0.0	0.00	0.00
%CV overall error	10.4	6.11	6.59

Abbreviation: %CV, coefficient of variation.

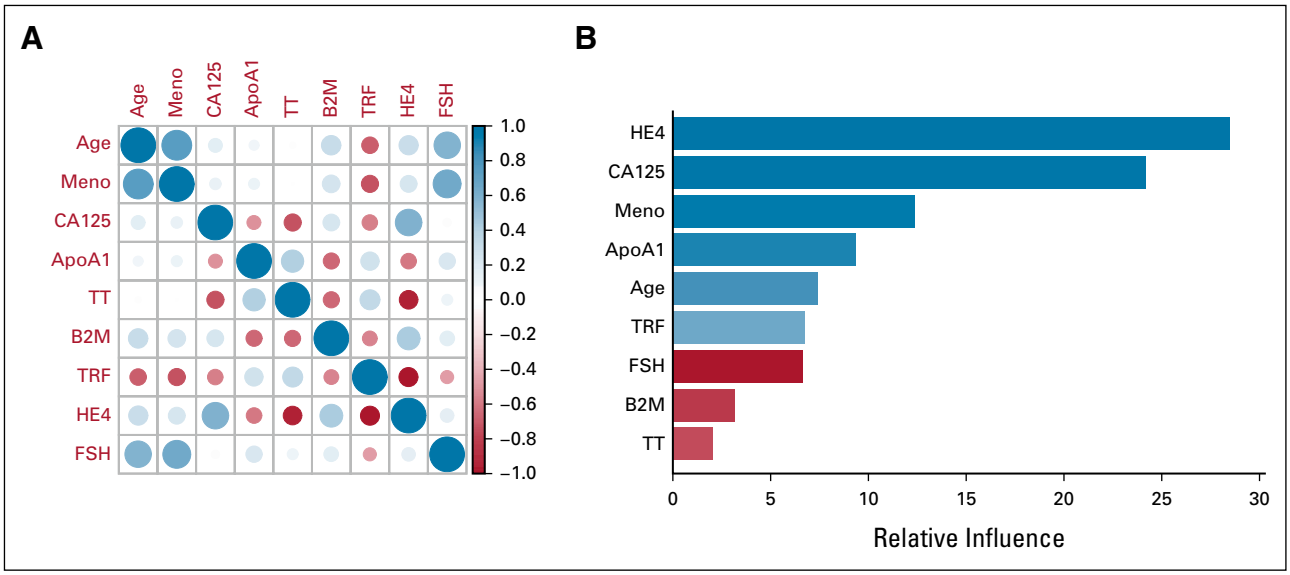

FIG A1. (A) Correlation matrix of the features used in the MIA3G algorithm. (B) Variable importance analysis of the features used in the MIA3G algorithm. ApoA1, apolipoprotein A1; B2M, beta-2 microglobulin; CA125, cancer antigen 125; FSH, follicle-stimulating hormone; HE4, human epididymis protein 4; Meno, menopausal status; TRF, transferrin; TT, transthyretin.