Medical Research Archives

European Society of Medicine

OPEN ACCESS

Published: November 30, 2023

Citation: Linder, K., et al., 2023. Prediction of Ovarian Cancer with Deep Machine Learning and Alternative Splicing. Medical Research Archives, [online] 11(11). https://doi.org/10.18103/mra.v11i11.4602

DOI: https://doi.org/10.18103/mra.v11i11.4602

ISSN: 2375-1924

RESEARCH ARTICLE

Prediction of Ovarian Cancer with Deep Machine Learning and Alternative Splicing

Katharine Linder[1], Rachel Watson[1], Keely Ulmer[1], David Bender[1], Michael J Goodheart[1], Eric Devor[1], Jesus Gonzalez Bosquet[1*]

[1]University of Iowa Hospitals and Clinics Department of Obstetrics and Gynecology, Iowa City, Iowa.

[*]jesus-gonzalezbosquet@uiowa.edu

ABSTRACT

Objective: Early detection of ovarian cancer could lead to improved survival rates, however no method has reliably been able to predict ovarian cancer. The aim of this study is to determine if processing alternative splicing data from high grade serous ovarian cancer patients using machine learning analytics will discriminate high grade serous ovarian cancer from normal fallopian tube samples. The ultimate goal would be to have a model that can predict high grade serous ovarian cancer with a blood test.

Methods: This is a case-control study of patients with confirmed high grade serous ovarian cancer and those undergoing salpingectomy for benign indications. RNA-sequencing was performed on all samples. RNA-sequence data was then put into Deep-learning augmented RNA-seq analysis of transcript splicing software suite. Deep-learning augmented RNA-seq analysis of transcript splicing created a model of differential alternative splicing aimed to discriminate between high grade serous ovarian cancer and normal fallopian tube. DEXSeq analysis was used to determine exon-based expression. Initial results with both analytics were then modelled with multivariate lasso regression to create prediction models (performance determined by area under the curve and 95% CI). Models created were the validated using The Cancer Genome Atlas data sets.

Results: One hundred and twelve high grade serous ovarian cancer and 12 benign samples were successfully sequenced. Deep-learning augmented RNA-sequencing analysis of transcript splicing identified 998 unique differentially expressed exons between high grade serous ovarian cancer and controls. Multivariate lasso regression analysis identified several exons that predicted high grade serous ovarian cancer with high performance. Specifically, ENSG00000182512:E001 from gene GLRX5 was highly predictive of high grade serous ovarian cancer with an area under the curve of 100%.

Conclusions: Application of machine learning analytics to exon differential expression, most likely due to alternative splicing, predicted high grade serous ovarian cancer with high performance. These results were validated in an independent dataset of cases and controls. Differential exon expression from cell-free RNA potentially could be used for early diagnosis of high grade serous ovarian cancer.

Keywords: ovarian cancer, alternative splicing, machine learning.

Introduction

There are estimated to be less than 20,000 new cases of ovarian cancer in the United States in 2022, but it is the fifth leading cause of cancer deaths in women[1]. Of the subtypes of ovarian cancer, high-grade serous carcinoma (HGSC) is the most common and encompasses cancers originating from the fallopian tube, ovary, and peritoneum. Ovarian cancer is most frequently diagnosed at advanced stages, and the 5-year survival rate for patients with ovarian cancer is less than 50%. However, when diagnosed at an early stage it is greater than 90%[2]. Early detection and treatment of ovarian cancer has been shown to decrease mortality, however, there is a lack of screening methods to achieve early detection[3].

Patients with ovarian cancer will typically present with a pelvic mass, however, the differential diagnosis for pelvic mass is broad and preoperative diagnosis of pelvic masses can be difficult. While cancer antigen 125 (CA-125) and transvaginal ultrasound (TVUS) are critical for the initial assessment of adnexal mass, each lacks sensitivity and specificity to be utilized as standard screening tests. CA-125 and TVUS have been utilized in large studies but failed to show a significant mortality benefit and illustrated the potential for harm for patients with benign masses undergoing unnecessary surgical intervention[4,5]. Studies have shown that patients with gynecologic cancers have better outcomes when treated by gynecologic oncologists[6–8]. Improved screening methods would lead to better healthcare utilization and patient outcomes by assessing which patients with adnexal masses should be evaluated by gynecologic oncologists.

There is emerging evidence for utilizing cell free DNA (cfDNA) and circulating tumor DNA (ctDNA) in a patient's plasma[9]. Circulating tumor DNA refers to the tumor DNA that is distinct from the patient DNA present in the plasma. This has become known as a "liquid biopsy." Studies have shown the potential for cfDNA and ctDNA to be utilized in the diagnosis and treatment strategies for several cancers including pancreatic, colorectal, lung, and breast[10–13]. We sought to create a machine learning (ML) prediction model analyzing RNA sequencing that could be used as a "liquid biopsy" and could diagnose HGSC patients at earlier stages.

Deep ML is a means of utilizing artificial intelligence to enhance our ability to process large amounts of data and make predictive models. Deep ML models use algorithms to analyze patterns in data sets, then using these patterns the model trains itself to make predictions on new data sets[14]. It has been used in oncology with varied applications, such as to classify disease genetic variations in genomes and identify RNA sequences[15,16]. A prior study developed Deep-Learning Augmented RNA-seq analysis of Transcript Splicing, or DARTS. This program involves a "deep neural network model that predicts differential alternative splicing between two conditions based on exon-specific sequence features and sample specific regulatory features; and a Bayesian Hypothesis (BHT) statistical model that infers differential alternative splicing by integrating empirical evidence in a specific RNA-seq dataset with prior probability of differential alternative splicing[17]." We will use DARTS, a ML based method, to create a prediction model that could detect HGSC in blood.

DEXseq is another method that has been used to test for differences in the usage of exons in RNA sequencing samples and, consequently, alternative splicing in biological samples. DEXseq has the ability to control for false positives by taking into account biological variability[18]. We will also use this method to assess differential exon expression to compare and complement DARTS method.

We utilized a large biobank of ovarian cancer specimens and normal fallopian tubes at our institution to analyze RNA splicing. We hypothesize that alternative splicing analyzed with ML will discriminate HGSC from normal fallopian tube samples. Our objective is to create a model to identify HGSC using two different methods for differential exon expression and alternative splicing, deep ML framework (DARTS) and DEXSeq methods. These models were further validated with independent datasets, TCGA, and in ML analytical platforms.

Methods

We performed a single-institution retrospective case-control study using HGSC tumor specimens obtained at the time of cytoreductive surgery and benign fallopian tube specimens. RNA was isolated from all specimens and RNA sequencing (RNA-seq) was performed.

SPECIMEN ACQUISITION

High grade serous ovarian cancer tissue samples were obtained from the Department of Obstetrics and Gynecology Gynecologic Oncology Biobank (IRB, ID no. 200209010), part of the Women's Health Tissue Repository (WHTR, IRB, ID no. 201804817) at the University of Iowa (UI). All tissues archived in the Gynecologic Oncology Biobank (herein termed Biobank) were originally obtained from adult patients under informed consent in accordance with University of Iowa Institutional Review Board (IRB) guidelines. Tumor samples were collected, reviewed by a board-certified pathologist, flash-frozen, and then the diagnosis was confirmed in paraffin at the time of initial surgery. All experimental protocols were approved by the University of Iowa Biomedical IRB-01.

Fallopian tube samples from patients undergoing gynecologic procedures were obtained from patients with no family history of cancer beside squamous cell carcinoma of the skin and who were undergoing salpingectomy for benign indications. (IRB, ID no. 201202714). These samples were similarly obtained from adult patients under informed consent in accordance with the University of Iowa IRB guidelines. RNA from both the fallopian tubes and HGSC specimens had already been extracted and purified in a previous study.

Two-hundred and fifty-three patients with advanced (stage II and IV) or recurrent HGSC were identified from the database. Of these 193 had available flash frozen tissue for RNA isolation. One-hundred and twelve of the samples had good quality RNA successfully processed for RNA-seq. Of the 20 benign fallopian tube samples obtained, 12 patients had good quality RNA that was successfully processed for RN-seq (**Figure 1A**). All samples were obtained and processed for previous studies[19].

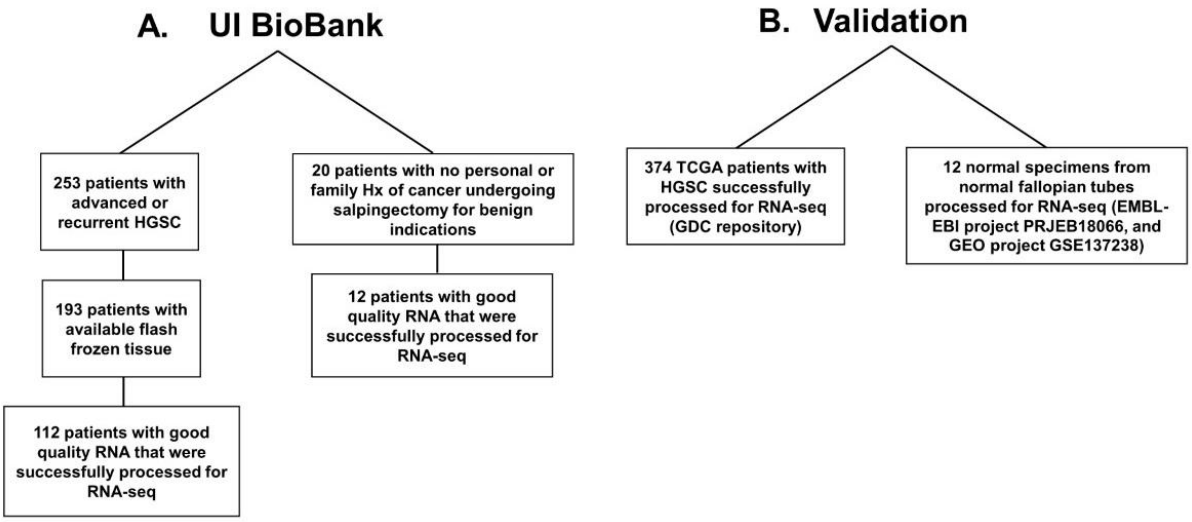

Figure 1: Flow chart of specimens used for analysis. A. Specimens processed from UI Biobank: 112 cases and 12 controls (normal fallopian tubes). B. Publicly available samples for validation: TCGA HSC specimens, and normal fallopian tubes from EMBL-EBI repository (project number PRJEB18066) and GEO accession number GSE137238.

RIBONUCLEIC ACID SEQUENCING

Ribonucleic acid was isolated from HGSC and benign fallopian tube samples. Methods for RNA sequencing and processing have been described elsewhere[19].Total cellular RNA was extracted from tissue with the mirVana (Thermo Fisher, Waltham, USA) RNA purification kit. The RNA yield and quality were assessed with Trinean Dropsense 16 spectrophotometer and Agilent Model 2100 bioanalyzer. RNA quality was determined to be adequate if the sample had an RNA integrity number (RIN) of 7.0 or greater. Samples that were of adequate quality were then sequenced. 500ng of RNA was quantified by Qubit measurement (Thermo Fisher). RNA was then converted to cDNA and ligated to sequencing adaptors with Illumina TriSeq stranded total RNA library preparation (Illumina, San Diego, CA, USA). Sequencing was then carried out on the Illumina HiSeq 4000 genome sequencing platform using 150 bp paired-end SBS chemistry. All sequencing was performed in the Genome Facility of the University of Iowa Institute of Human Genetics (IIHG).

DEEP-LEARNING AUGMENTED RNA-SEQ ANALYSIS OF TRANSCRIPT SPLICING ANALYSIS

Deep-learning augmented RNA-seq analysis of transcript splicing (DARTS) is a method for predicting alternative splicing patterns[17]. It uses a bayesian hypothesis testing (BHT) model paired with a deep neural network (DNN) to predict differential alternative splicing. The DARTS BHT initially analyzes large-scale RNA-seq data to generate training labels for differential or unchanged splicing events. This is used to train the DNN. Then the trained DARTS DNN was used to predict alternative splicing and differential exon expression in our data set. The prediction, along with observed RNA-seq read counts, are incorporated as an informative prior by the DARTS BHT to perform deep learning augmented splicing analysis.

DEXSEQ ANALYSIS

DEXSeq analysis was used to determine exon-based expression[18]. BAM files from RNAseq were used to determine expression of single

exons. The DESeq2 R package was used to normalize and transform data to determine the differences between HGSC samples and benign fallopian tube samples[20] (**Supplementary Figure S1 and S2**). The differential expression analysis in DESeq2 package uses a generalized linear model to assess difference in log2-transformed counts. A p-value (Wald test p-value) adjusted with false discovery rate (FDR) was considered significant when < 0.0001 to account for multiple comparisons[21].

CREATION OF PREDICTION MODELS OF HIGH GRADE SEROUS OVARIAN CANCER) WITH MULTIVARIATE MODELS

Significant variables from the univariate analysis - ANOVA for the DEXSeq analysis and DARTS analysis - were then incorporated into two different multivariate lasso regression prediction models of ovarian cancer (HGSC), one for each method. Multivariate prediction models were fit with lasso (least absolute shrinkage and selection operator) as implemented in the glmnet R package[22], and detailed previously[21]. Performances of prediction models were measured by the area under the receiver operating characteristics curve (AUC) and their 95% confidence interval (CI), estimated with 1,000 replicates of ten-fold cross-validation to avoid over-fitting. Bias-corrected and accelerated bootstrap CI's were computed for each model. AUC of 0.5 indicates no predictive ability of a model and 1.0 represents perfect predictive performance.

VALIDATION OF PREDICTIVE MODELS WITH THE CANCER GENOME ATLAS DATABASE AND MACHINE LEARNING ANALYTICS

High grade serous ovarian cancer data from The Cancer Genome Atlas (TCGA) was used to validate created predictive models[23]. We included publicly available controls from EMBL-EBI project PRJEB18066 and GEO project GSE137238 databases (**Figure 1B**). After permission was granted to access controlled data by the Genomic Data Commons (GDC) Data Portal (dbGaP# 29868), TCGA HGSC BAM files from RNA-seq experiments aligned to the human reference genome (version hg38) with the STAR suite were downloaded in their original format. DEXSeq analysis to determine exon-based expression was performed as described previously for both, TCGA cases and GEO and EMBL-EBI controls. Then selected features used in the multivariate prediction lasso model were extracted also from TCGA single exon expression dataset and used for prediction model validation of both, DARTS and DEXSeq analyses.

Machine learning prediction models with TensorFlow platform were performed with those features more informative for prediction of HGSC in previous analysis. Initial validation of prediction models with ML included all selected exons from initial univariate and multivariate analyses for both DARTS and DEXSeq. Then we used ML platforms to validate models performed with those exons selected in TCGA validation analysis. The final goal was to identify and validate the simplest, most accurate and robust model that could predict HGSC.

Results

We extracted count expression information from 645,243 exons from over 60,000 different transcripts in our database, including genes, long noncoding RNAs (lncRNAs), MIRs and other transcription units. Differential

exon expression was performed with both DARTS and DEXSeq analysis (**Figure 2**). Deep-learning augmented RNA-seq analysis of transcript splicing identified 998 unique differentially expressed exons between HGSC and controls (**Figure 2A, upper panel**). A prediction model with all these exons had an AUC of 91%. Differential exon expression with DEXSeq, identified 1,399 significant exons (p<0,001).

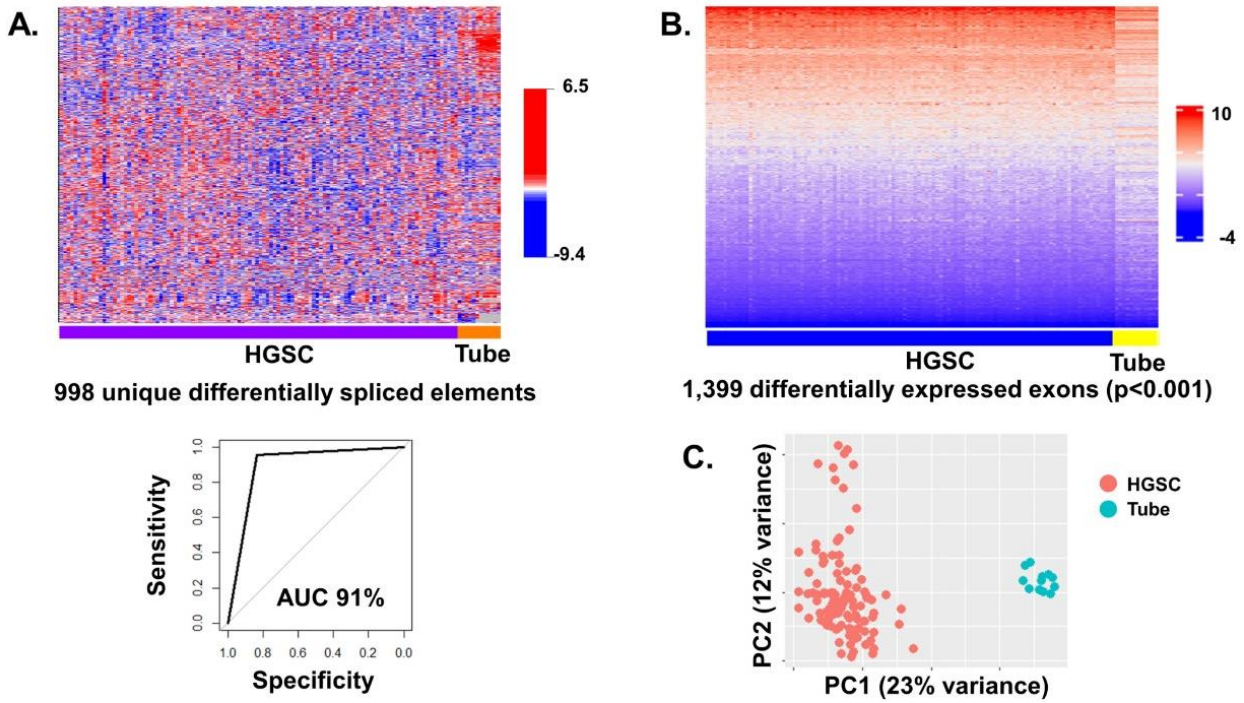

Figure 2: Analysis of differential exon expression and alternative splicing by different methods: **A.** *DARTS*: first a deep neural network (DNN) model predicts differential alternative splicing based on exon-specific sequence features and sample-specific features (upper panel); then, a Bayesian hypothesis testing (BHT) statistical model infers differential alternative splicing by integrating empirical evidence in a specific RNA-seq dataset with prior probability of differential alternative splicing. Prediction model represented by the ROC curve (lower panel). **B.** Differential exon expression by *DEXSeq* analysis. Heatmap with resultant exons. **C.** Principal component (PC) analysis of differential exon expression with *DEXSeq*. The figure represents a plot of the two main PCs.

CREATION OF MULTIVARIATE PREDICTION MODELS OF HIGH GRADE SEROUS OVARIAN CANCER

A multivariate lasso regression prediction model using all features selected in the univariate analysis after DEXSeq assessment (N=1,399) identified a unique exon, ENSG00000182512:E001 (within gene *GLRX5*) that predicted HGSC with and AUC of 100% (**Figure 3A**, upper panel). The multivariate lasso regression using the results from the DARTS analysis identified three exons: ENSG00000050130:E013 (within gene *JKAMP*), ENSG00000135597:E014 (in gene *REPS1*), and ENSG00000175061:E013 (in lncRNA *FAM211A-AS1*, or *SNHG29*), which predicted HGSC with an AUC of 100% (**Figure 3A**, lower panel). When assessing the relative expression between of these 4 exons between HGSC and tube samples, only ENSG00000182512:E001 increased risk for HGSC, the other three were decreased, and protective of HGSC (**Figure 3B**).

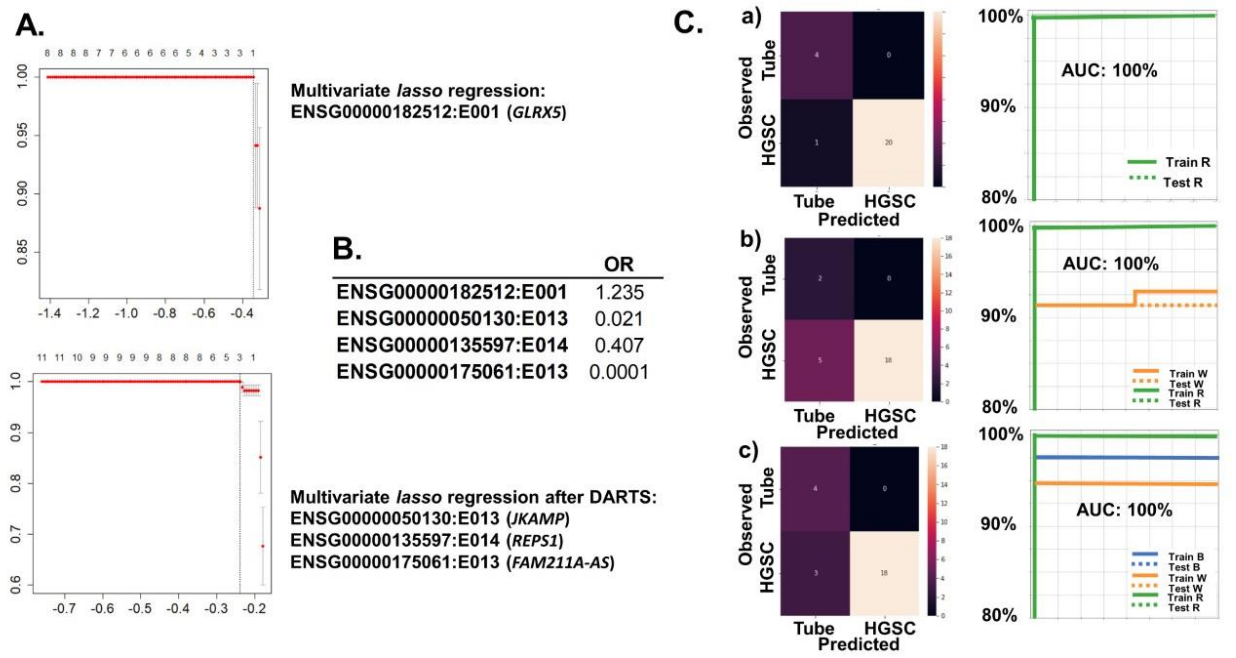

Figure 3: Multivariate lasso regression analysis and machine learning validation with TensorFlow.

A. Multivariate lasso regression analysis to predict HGSC with differential exon expression: In the upper panel, multivariate lasso regression prediction model using results from univariate regression analysis after DEXSeq ($p<0.001$): one differentially expressed exon, ENSG00000182512:E001 (*GLRX5*) predicted HGSC with and AUC of 100%; the inferior panel shows the multivariate lasso regression prediction model using results from the machine learning augmented method DARTS: three exons, ENSG00000050130:E013 (*JKAMP*), ENSG00000135597:E014 (*REPS1*), and ENSG00000175061:E013 (*FAM211A-AS*), predicted HGSC with an AUC of 100%. **B.** Exons odds ratio expression with respect to normal Fallopian tube: only ENSG00000182512:E001 (*GLRX5*) expression was elevated in HGSC vs tube; the expression of the other 3 exons were lower in HGSC samples.**C.** Validation of HGSC prediction models with machine learning using TensorFlow: **a)** Model using all differentially expressed exon in DEXSeq univariate analysis (N=1,399): the left panel shows the confusion matrix representing the observed versus the predicted values; the right panel represents the ROC graphic including a model accounting for unbalanced samples: Train R: results of unbalanced (or re-sampling) model training; Test R: results of re-sampling model testing. **b)** Model using all 998 unique differentially spliced elements in the DARTS analysis (N=998): the left panel shows the confusion matrix representing the observed versus the predicted values; the right panel represents the ROC graphic including: 1) models accounting for weights of the outcome: Train W: results of weighted model training; Test W: results of weighted model testing; 2) models accounting for unbalanced samples: Train R: results of unbalanced (or re-sampling) model training; Test R: results of re-sampling model testing. **c)** Model using all 4 exons found to be significant in both multivariate lasso analyses: the left panel shows the confusion matrix representing the observed versus the predicted values; the right panel represents the ROC graphic including: 1) basic model: Train B: results of basic model training; Test B: results of basic model testing; 2) models accounting for weights of the outcome: Train W: results of weighted model training; Test W: results of weighted model testing; 3) models accounting for unbalanced samples: Train R: results of unbalanced (or re-sampling) model training; Test R: results of re-sampling model testing.

VALIDATION OF PREDICTION MODELS

Initially, these multivariate prediction models were validated in a different ML analytical platform using *TensorFlow*. First, we validated the models including features selected after the univariate analysis with DEXSeq and the initial DARTS assessment. The performance of prediction models with 1,399 exon from the DEXSeq analysis (**Figure 3C.a**) and 998 exons selected by DARTS (**Figure 3C.b**) had excellent performances, with AUCs of 100%. A third model with only the 4 resulting exons

after both multivariate lasso regressions (for DEXSeq and DARTS results), resulted in another model with excellent performance also, AUC of 100% (**Figure 3C.c**).

Then, these models were validated in an independent dataset. After downloading, homogenizing, determining exon expression, we normalized and log2 transformed TCGA and control data from EMBL-EBI and GEO databases. Out of the initial 1,399 exons selected after univariate analyses with DEXSeq, 1,171 were also detected in TCGA+controls databases. A new multivariate lasso regression was done with those 1,171 exons in UI dataset, and then the model was applied to TCGA+controls dataset using *pROC* (a R package) in data (**Figure 4A**). The same unique exon, ENSG00000182512 :E001 (in gene *GLRX5*) predicted HGSC in UI data with an AUC of 100%, and when the model was applied to TCGA data, the performance of the validated model was excellent, with an AUC of 97% (**Figure 4C**, upper panel).

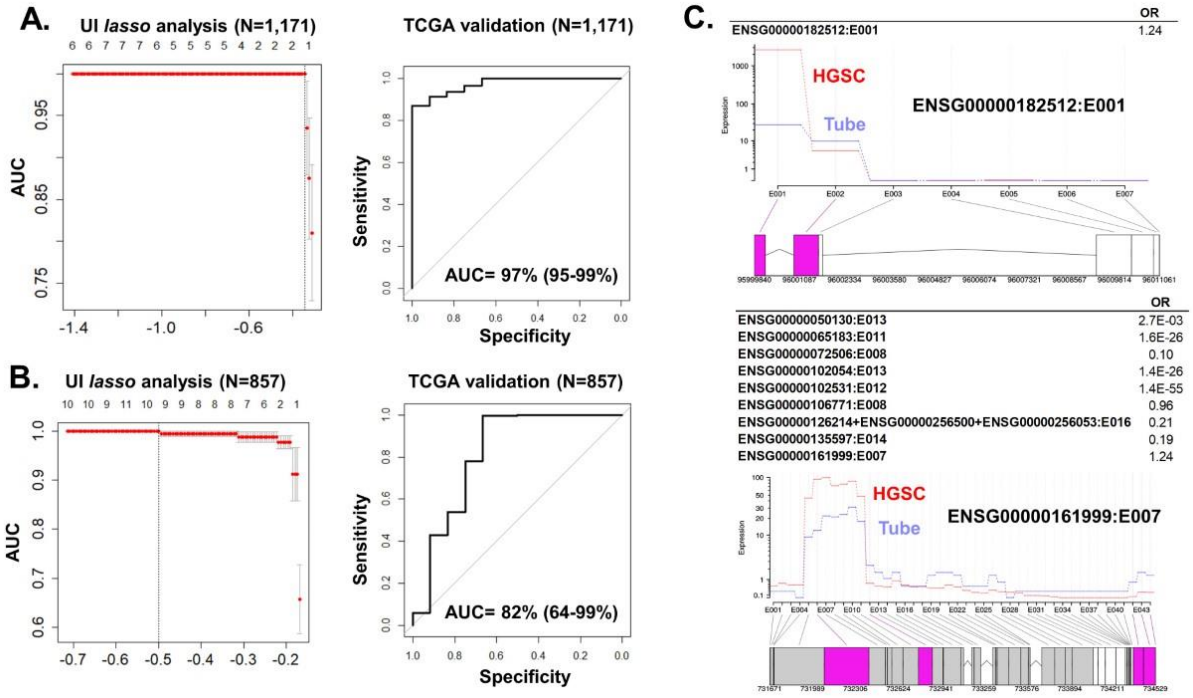

Figure 4: Validation of differential exon expression with TCGA data.

A. Validation of HGSC prediction with differential exon expression of DEXSeq analysis: the left panel, shows the prediction model using exon expression significant in the UI cohort, that is also present in TCGA cohort (N=1,171): again, one differentially expressed exon, ENSG00000182512:E001 (*GLRX5*) predicted HGSC with and AUC of 100%. The right panel shows the validation of the prediction model in TCGA data, with an excellent performance of 97% (CI, 95-99%).

B. Validation of HGSC prediction with differential exon expression regression prediction model using results from the machine learning augmented method DARTS: the left panel, shows the prediction model using exon expression also present in TCGA cohort (N=857): 9 differentially expressed exons predicted HGSC with and AUC of 100%. The right panel shows the validation of the prediction model in TCGA data, with a very good performance of 82% (CI, 64-99%).

C. Detailed level of exon expression in 2 of the exons. Both, ENSG00000182512:E001 and ENSG00000161999:E007, had higher expression in HGSC samples (red) than in tubal samples (blue). Comparison of other exons in the same genes are also represented.

Out of the initial 998 exons selected with DARTs, 857 exons were also detected in TCGA+controls dataset. A new multivariate lasso regression with these 857 exons in UI data, resulted in a prediction model of HGSC with an AUC of 100%. When this model was applied to TCGA+controls data, the resulting validated model also had very good performance with an AUC of 82% (**Figure 4B**). This TCGA validation model had 9 exons that were informative for HGSC classification (**Figure 4C**, mid and lower panel). Examples of alternative splicing based on expression of two exons selected in the multivariate analyses are represented in **Figure 4C** panel.

SIMPLIFIED PREDICTION MODEL OF HIGH GRADE SEROUS OVARIAN CANCER

With the goal to create a simplified model that is robust (across databases from different populations) and validated (across different analytical platforms and independent databases) we designed a new model with the most informative exons in previous models, with DEXSeq and DARTS selection, and after TCGA validation. For this new model we used the 10 exons more informative in predicting HGSC in previous models. The resulting model in UI data selected the same unique exon, ENSG00000182512:E001 (*GLRX5*) as predictor of HGSC with an AUC of 100% (**Figure 5A**, upper panel). When this model was applied to TCGA dataset, the validated model had a performance of 97% (**Figure 5A**, lower panel). Finally, we validated these simplified models in a new ML platform. First, we tested a model with the 10 selected exons in UI data, with excellent performance (**Figure 5B.a**). Then, we tested the model with the singular most predictive exon, ENSG00000182512:E001, in UI data also with excellent performance (**Figure 5B.b**). Lastly, we validated the model with only one exon in TCGA data, also with excellent performance of 97% (**Figure 5B.c**).

.

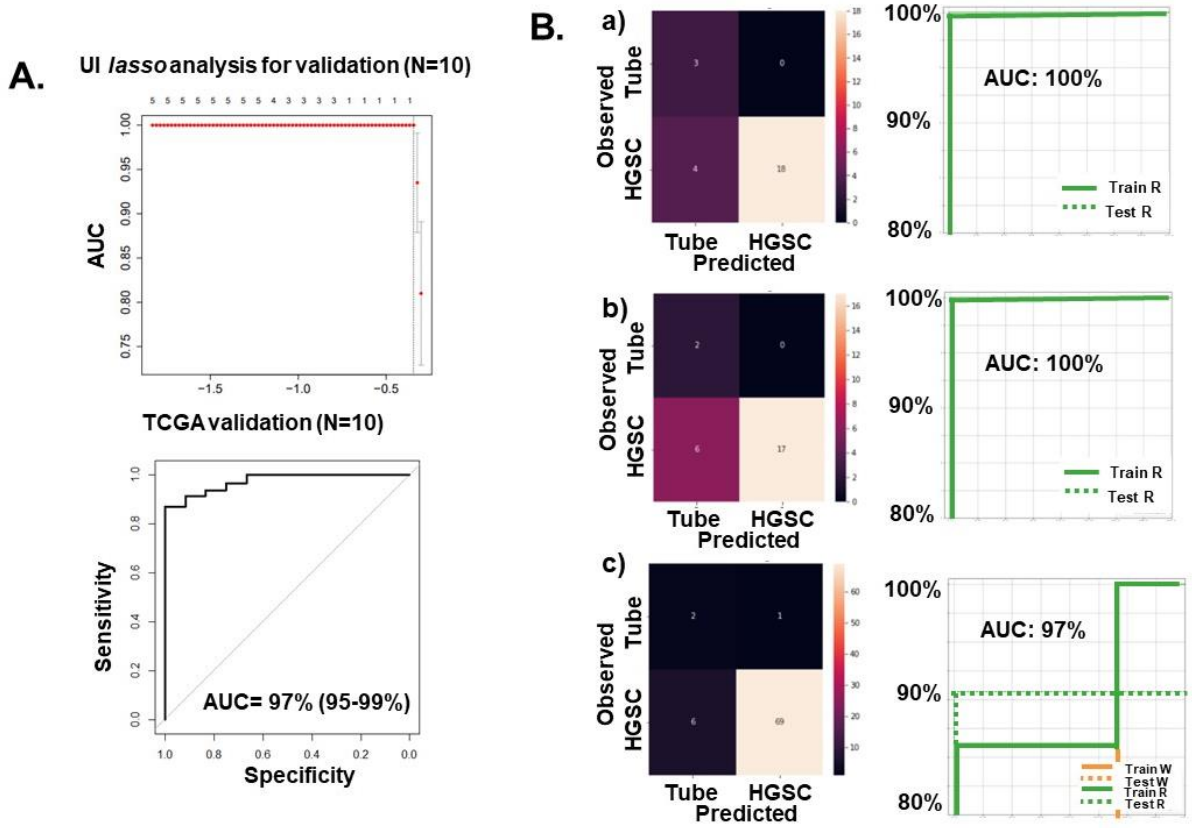

Figure 5: Validation of simplified prediction models of HGSC.

A. Validation of HGSC prediction with differential expression of the most relevant 10 exons in TCGA: the upper panel, shows the prediction model using the 10 more relevant exon expression (see Figure 4C) in the UI cohort: again, one exon, ENSG00000182512:E001 (*GLRX5*) predicted HGSC with and AUC of 100%. The lower panel shows the validation of the prediction model in TCGA data, with an excellent performance of 97% (CI, 95-99%).

B. Validation of HGSC prediction models with machine learning using TensorFlow: **a)** Model using the 10 more relevant exon expressions (see Figure 4C) in the UI cohort: the left panel shows the confusion matrix representing the observed versus the predicted values; the right panel represents the ROC graphic including models accounting for unbalanced samples: Train R: results of unbalanced (or re-sampling) model training; Test R: results of re-sampling model testing.

b) Model using only the more relevant exon expression (ENSG00000182512:E001) in the UI cohort: the left panel shows the confusion matrix representing the observed versus the predicted values; the right panel represents the ROC graphic with a model accounting for unbalanced samples: Train R: results of unbalanced (or re-sampling) model training; Test R: results of re-sampling model testing.

C. Validation of HGSC prediction models with machine learning using TensorFlow in TCGA data: Model using only the more relevant exon expression (ENSG00000182512:E001) in the UI cohort: the left panel shows the confusion matrix representing the observed versus the predicted values; the right panel represents the ROC graphic including: 1) model accounting for weights of the outcome: Train W: results of weighted model training; Test W: results of weighted model testing; 2) model accounting for unbalanced samples: Train R: results of unbalanced (or re-sampling) model training; Test R: results of re-sampling model testing.

Discussion

The aim of this study was to create a validated, robust model to identify HGSC from benign fallopian tube tissue using two different methods for differential exon expression and alternative splicing, deep machine learning framework (DARTS) and DEXSeq methods. There are currently no non-invasive methods available for the diagnosis of HGSC. TVUS and CA-125 have been used a surrogate marker for predicting malignant potential of adnexal masses, however, these lack specificity and sensitivity for screening tests[4,5]. The model created in this study could be used to detect ovarian cancer (HGSC) with blood tests, by detecting differences in

products (exons) originating from circulating tumor cells[24]. Survival outcomes for patients with ovarian cancer differ greatly depending on the stage at time of diagnosis[2]. Ultimately, the findings of this study could lead to the development of methods to diagnose HGSC early and non-invasively, which could improve survival outcomes. The simplified prediction model performed with a multivariate lasso regression analysis, identified several exons that predicted HGSC with high performance. Multivariate lasso regression analysis identified several exons that predicted HGSC with high performance. Specifically, ENSG00000182512:E001 from gene GLRX5 was highly predictive of HGSC with an AUC of 100%. Results were then successfully validated in public datasets TGCA and EMBL-EBI with excellent performance.

Alternative splicing has been identified as a possible driving force in several complex human diseases. Reble, Dineen, and Barr previously reviewed several studies that showed numerous genes with alternative splicing that are associated with several psychiatric conditions, including schizophrenia, bipolar disorder, and autism spectrum disorder[25]. Both type 1 and type 2 diabetes have been shown to have higher rates of alternative splicing genes that play a role in the development of these conditions[26,27]. Several cardiovascular diseases such as hypertension, atherosclerosis, and heart failure have been associated with alternative splicing in L-type calcium channels as previously reviewed by Hu, Liang, and Soong[28]. Additionally, obesity-related hypertension also has been associated with alternative splicing in several different genes[29]. Alternative splicing has been found to have prediction capabilities for prognostic and progression factors in several different types of cancer, including triple negative breast, gastric, testicular, melanoma, and hepatocellular[30–33], as well as ovarian cancer.

Studies have identified alternative splicing patterns associated with ovarian cancer[34] and comparison with normal samples have identified unique exon-exon junctions more common in ovarian cancer[35]. There have been studies that showed promising results for predicting overall survival in ovarian cancer patients based on alternative splicing patterns[36], but few have looked for alternative splicing signatures that may be used to diagnose ovarian cancer. The predictive model in our study identified an exon from alternative splicing from the gene Glutaredoxin 5 (GLRX5). This gene codes for a protein localized to the mitochondria and is involved in biogenesis of ion-sulfur clusters[37]. Expression of GLRX5 has been shown to be altered in hepatocellular carcinoma[38]. Other studies have identified inhibition of the protein has been associated with cisplatin-resistant head and neck cancers[39], however no studies we are aware of have linked GLRX5 and ovarian cancer. Furthermore, the exon described has not been associated with ovarian cancer or described previously. In fact, no studies that we are aware of have linked alternative splicing of GLRX5 to outcomes in any cancer. The identification of ENSG00000182512:E001 from gene GLRX5 has the potential for early diagnosis of ovarian cancer and extrapolating findings from other studies identifying GLRX5 could also have prognostic or therapeutic implications for ovarian cancer.

The strengths of this study are the use of data from a clinically well annotated database of patients with HGSC with comprehensive genomic information and an adequate number of controls. Additionally, prediction models were validated in an independent comprehensive dataset (TCGA). Further, data were analyzed with two separate analytical platforms. The limitations of this study are that it is retrospective, and therefore is at risk of bias. The UI database is limited by minimal diversity in the sample given the demographics of Iowa, with the general population made up of 90.1% white individuals[40]. Future studies using these models prospectively and in larger data sets are needed to test their clinical utility.

Conclusion

Our study created a validated model for identifying HGSC by utilizing deep machine learning framework (DARTS) and DEXseq to identify differences in exon expression and alternative splicing in tumor cells when compared to benign fallopian tube cells. In the future, this model could be utilized to differentiate tumor cells in blood samples as a means for early and non-invasive diagnosis of HGSC, which could ultimately lead to a reduction in mortality.

Conflict of Interest Statement:

The authors have no conflicts of interest to disclose.

Acknowledgement Statement:

Funding: This work was supported in part by the NIH grant R01 CA99908 and R01 CA184101 to Kimberly K. Leslie, and the basic research fund from the Department of Obstetrics & Gynecology at the University of Iowa. Also, was supported in part by the American Association of Obstetricians and Gynecologists Foundation (AAOGF) Bridge Funding Award.

Funding Statement:

None

References:

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin*. 2022;72(1):7-33. doi:10.3322/caac.21708

2. SEER Cancer Statistics Review, 1975-2016. SEER. Accessed March 20, 2023. https://seer.cancer.gov/csr/1975_2016/index.html

3. Henderson JT, Webber EM, Sawaya GF. Screening for Ovarian Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA*. 2018;319(6):595-606. doi:10.1001/jama.2017.21421

4. Pinsky PF, Yu K, Kramer BS, et al. Extended mortality results for ovarian cancer screening in the PLCO trial with median 15years follow-up. *Gynecol Oncol*. 2016; 143(2):270-275. doi:10.1016/j.ygyno.2016.08.334

5. Buys SS, Partridge E, Black A, et al. Effect of screening on ovarian cancer mortality: the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Randomized Controlled Trial. *JAMA*. 2011; 305(22):2295-2303. doi:10.1001/jama.2011.766

6. Junor EJ, Hole DJ, McNulty L, Mason M, Young J. Specialist gynaecologists and survival outcome in ovarian cancer: a Scottish national study of 1866 patients. *Br J Obstet Gynaecol*. 1999; 106(11):1130-1136. doi:10.1111/j.1471-0528.1999.tb08137.x

7. McGowan L, Lesher LP, Norris HJ, Barnett M. Misstaging of ovarian cancer. *Obstet Gynecol*. 1985; 65(4):568-572.

8. Im SS, Gordon AN, Buttin BM, et al. Validation of referral guidelines for women with pelvic masses. *Obstet Gynecol*. 2005; 105(1):35-41. doi:10.1097/01.AOG.0000149159.69560.ef

9. Bettegowda C, Sausen M, Leary RJ, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med*. 2014; 6(224):224ra24. doi:10.1126/scitranslmed.3007094

10. Mellby LD, Nyberg AP, Johansen JS, et al. Serum Biomarker Signature-Based Liquid Biopsy for Diagnosis of Early-Stage Pancreatic Cancer. *J Clin Oncol Off J Am Soc Clin Oncol*. 2018;36(28):2887-2894. doi:10.1200/JCO.2017.77.6658

11. Abbosh C, Birkbak NJ, Wilson GA, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature*. 2017; 545(7655):446-451. doi:10.1038/nature22364

12. Phallen J, Sausen M, Adleff V, et al. Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med*. 2017; 9(403):eaan2415. doi:10.1126/scitranslmed.aan2415

13. Cheng X, Zhang L, Chen Y, Qing C. Circulating cell-free DNA and circulating tumor cells, the "liquid biopsies" in ovarian cancer. *J Ovarian Res*. 2017; 10(1):75. doi:10.1186/s13048-017-0369-5

14. Sultan AS, Elgharib MA, Tavares T, Jessri M, Basile JR. The use of artificial intelligence, machine learning and deep learning in oncologic histopathology. *J Oral Pathol Med Off Publ Int Assoc Oral Pathol Am Acad Oral Pathol*. 2020; 49(9):849-856. doi:10.1111/jop.13042

15. Leung MKK, Xiong HY, Lee LJ, Frey BJ. Deep learning of the tissue-regulated splicing code. *Bioinformatics*. 2014; 30(12):i121-i129. doi:10.1093/bioinformatics/btu277

16. Xiong HY, Alipanahi B, Lee LJ, et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science*.

2015;347(6218):1254806. doi:10.1126/science.1254806

17. Zhang Z, Pan Z, Ying Y, et al. Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat Methods*. 2019;16(4):307-310. doi:10.1038/s41592-019-0351-9

18. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res*. 2012; 22(10):2008-2017. doi:10.1101/gr.133744.111

19. Gonzalez-Bosquet J, Cardillo ND, Reyes HD, et al. Using Genomic Variation to Distinguish Ovarian High-Grade Serous Carcinoma from Benign Fallopian Tubes. *Int J Mol Sci*. 2022; 23(23):14814. doi:10.3390/ijms232314814

20. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010; 11(10):R106. doi:10.1186/gb-2010-11-10-r106

21. Gonzalez Bosquet J, Devor EJ, Newtson AM, et al. Creation and validation of models to predict response to primary treatment in serous ovarian cancer. *Sci Rep*. 2021; 11(1):5957. doi:10.1038/s41598-021-85256-9

22. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010; 33(1):1-22.

23. Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Softw*. 2008; 28:1-26. doi:10.18637/jss.v028.i05

24. Asante DB, Calapre L, Ziman M, Meniawy TM, Gray ES. Liquid biopsy in ovarian cancer using circulating tumor DNA and cells: Ready for prime time? *Cancer Lett*. 2020; 468:59-71. doi:10.1016/j.canlet.2019.10.014

25. Reble E, Dineen A, Barr CL. The contribution of alternative splicing to genetic risk for psychiatric disorders. *Genes Brain Behav*. 2018;17(3):e12430. doi:10.1111/gbb.12430

26. Dlamini Z, Mokoena F, Hull R. Abnormalities in alternative splicing in diabetes: therapeutic targets. *J Mol Endocrinol*. 2017; 59(2):R93-R107. doi: 10.1530/JME-17-0049

27. Rehman SU, Schallschmidt T, Rasche A, et al. Alternative exon splicing and differential expression in pancreatic islets reveals candidate genes and pathways implicated in early diabetes development. *Mamm Genome Off J Int Mamm Genome Soc*. 2021; 32(3):153-172. doi:10.1007/s00335-021-09869-1

28. Hu Z, Liang MC, Soong TW. Alternative Splicing of L-type CaV1.2 Calcium Channels: Implications in Cardiovascular Diseases. *Genes*. 2017; 8(12):344. doi:10.3390/genes8120344

29. Dlamini Z, Hull R, Makhafola TJ, Mbele M. Regulation of alternative splicing in obesity-induced hypertension. *Diabetes Metab Syndr Obes Targets Ther*. 2019; 12:1597-1615. doi:10.2147/DMSO.S188680

30. Yu S, Hu C, Liu L, et al. Comprehensive analysis and establishment of a prediction model of alternative splicing events reveal the prognostic predictor and immune microenvironment signatures in triple negative breast cancer. *J Transl Med*. 2020; 18(1):286. doi:10.1186/s12967-020-02454-1

31. Lou S, Zhang J, Zhai Z, et al. Development and validation of an individual alternative splicing prognostic signature in gastric cancer. *Aging*. 2021; 13(4):5824-5844. doi:10.18632/aging.202507

32. Zhang CJ, Li ZT, Shen KJ, Chen L, Xu DF, Gao Y. Characterization of progression-related alternative splicing events in testicular germ cell tumors. *Asian J Androl*. 2021; 23(3):259-265. doi:10.4103/aja.aja_30_20

33. Wan Q, Sang X, Jin L, Wang Z. Alternative Splicing Events as Indicators for the Prognosis of Uveal Melanoma. *Genes*. 2020; 11(2):227. doi:10.3390/genes11020227

34. Klinck R, Bramard A, Inkel L, et al. Multiple alternative splicing markers for ovarian cancer. *Cancer Res*. 2008; 68(3):657-663. doi:10.1158/0008-5472.CAN-07-2580

35. Kahles A, Lehmann KV, Toussaint NC, et al. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell*. 2018; 34(2):211-224.e6. doi:10.1016/j.ccell.2018.07.001

36. Zhu J, Chen Z, Yong L. Systematic profiling of alternative splicing signature reveals prognostic predictor for ovarian cancer. *Gynecol Oncol*. 2018; 148(2):368-374. doi:10.1016/j.ygyno.2017.11.028

37. Ogata FT, Branco V, Vale FF, Coppo L. Glutaredoxin: Discovery, redox defense and much more. *Redox Biol*. 2021;43:101975. doi:10.1016/j.redox.2021.101975

38. Mollbrink A, Jawad R, Vlamis-Gardikas A, et al. Expression of Thioredoxins and Glutaredoxins in Human Hepatocellular Carcinoma: Correlation to Cell Proliferation, Tumor Size and Metabolic Syndrome. *Int J Immunopathol Pharmacol*. 2014; 27(2):169-183. doi:10.1177/039463201402700204

39. Lee J, You JH, Shin D, Roh JL. Inhibition of Glutaredoxin 5 predisposes Cisplatin-resistant Head and Neck Cancer Cells to Ferroptosis. *Theranostics*. 2020; 10(17):7775-7786. doi:10.7150/thno.46903

40. U.S. Census Bureau QuickFacts: Iowa. Accessed March 20, 2023. https://www.census.gov/quickfacts/IA