




# RoBERTa-Assisted Outcome Prediction in Ovarian Cancer Cytoreductive Surgery Using Operative Notes

Cancer Control  
Volume 30: 1–11  
© The Author(s) 2023  
Article reuse guidelines:  
[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)  
DOI: 10.1177/10732748231209892  
[journals.sagepub.com/home/ccx](https://journals.sagepub.com/home/ccx)



Alexandros Laios, MD, PhD, PostDoc (Ox)<sup>1</sup> , Evangelos Kalampokis, PhD<sup>2</sup>,  
Marios Evangelos Mamalis, BA, MSc<sup>2</sup> , Constantine Tarabanis, MD<sup>3</sup>,  
David Nugent, MBChB, MRCOG, PhD<sup>1</sup>, Amudha Thangavelu, MBChB, MRCOG, MD<sup>1</sup>,  
Georgios Theophilou, MBBS, MRCOG, MD<sup>1</sup>, and Diederick De Jong, MBBCh, PhD, MSc<sup>1</sup> 

## Abstract

**Introduction:** Contemporary efforts to predict surgical outcomes focus on the associations between traditional discrete surgical risk factors. We aimed to determine whether natural language processing (NLP) of unstructured operative notes improves the prediction of residual disease in women with advanced epithelial ovarian cancer (EOC) following cytoreductive surgery.

**Methods:** Electronic Health Records were queried to identify women with advanced EOC including their operative notes. The Term Frequency – Inverse Document Frequency (TF-IDF) score was used to quantify the discrimination capacity of sequences of words (n-grams) regarding the existence of residual disease. We employed the state-of-the-art RoBERTa-based classifier to process unstructured surgical notes. Discrimination was measured using standard performance metrics. An XGBoost model was then trained on the same dataset using both discrete and engineered clinical features along with the probabilities outputted by the RoBERTa classifier.

**Results:** The cohort consisted of 555 cases of EOC cytoreduction performed by eight surgeons between January 2014 and December 2019. Discrete word clouds weighted by n-gram TF-IDF score difference between R0 and non-R0 resection were identified. The words ‘adherent’ and ‘miliary disease’ best discriminated between the two groups. The RoBERTa model reached high evaluation metrics (AUROC .86; AUPRC .87, precision, recall, and F1 score of .77 and accuracy of .81). Equally, it outperformed models that used discrete clinical and engineered features and outplayed the performance of other state-of-the-art NLP tools. When the probabilities from the RoBERTa classifier were combined with commonly used predictors in the XGBoost model, a marginal improvement in the overall model’s performance was observed (AUROC and AUPRC of .91, with all other metrics the same).

**Conclusion/Implications:** We applied a sui generis approach to extract information from the abundant textual surgical data and demonstrated how it can be effectively used for classification prediction, outperforming models relying on conventional structured data. State-of-art NLP applications in biomedical texts can improve modern EOC care.

<sup>1</sup>Department of Gynaecologic Oncology, ESGO Center of Excellence for Ovarian Cancer Surgery, St James’s University Hospital, Leeds, UK

<sup>2</sup>Information Systems Lab, Department of Business Administration, University of Macedonia, Thessaloniki, Greece

<sup>3</sup>Department of Internal Medicine, School of Medicine, New York University, New York, NY, USA

## Corresponding Author:

Alexandros Laios, Department of Gynaecologic Oncology, St James’s University Hospital, Beckett street, Leeds LS9 7TF, UK.

Email: [a.laios@nhs.net](mailto:a.laios@nhs.net)



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and

Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

## Keywords

epithelial ovarian cancer, complete cytoreduction, operative notes, natural language processing, machine learning, transfer learning, RoBERTa, explainable artificial intelligence

Received June 15, 2023. Received revised September 16, 2023. Accepted for publication September 29, 2023.

## Introduction

Contemporary efforts to predict surgical outcomes and postoperative complications usually focus on the associations between traditional surgical risk factors, including age or preoperative albumin.<sup>1,2</sup> In addition to risk factors in discrete data fields, we now have access to abundant textual data within digital medical records. In the era of healthcare digitalization, the increasing implementation of Electronic Health Records (EHRs) at UK Hospitals has created valuable data sources for clinical and translational research.<sup>3</sup> Although EHRs hold structured data, a large proportion of clinical notes are in narrative text format. It is estimated that unstructured data accounts for more than 80% of currently available healthcare data.<sup>4</sup> Reading note text and extracting information are resource intensive. Artificial intelligence (AI) has emerged as a potential solution for harnessing these data. More specifically, Natural Language Processing (NLP) is the AI discipline that focuses on extracting information from texts by converting narrative clinical notes into a structured format. The NLP methods have been shown to achieve remarkable results in such tasks using hundreds to thousands of clinical notes.<sup>5</sup> Their implementation has been promoted and accelerated during the COVID era.<sup>6</sup> Nevertheless, clinical research has been heavily affected by the underutilization of unstructured data from EHRs.<sup>7</sup>

Amongst the best NLP models employed to date, the Bidirectional Encoder Representations from Transformers (BERT) model was created by Google in 2018. Thanks to its architecture, it can extract information from texts by considering bidirectional contextual information.<sup>8</sup> BERT's advanced information extraction capacities when combined with the fact that traditional NLP methods such as Word2vec have shown promising results in classification tasks in clinical settings and<sup>9</sup> can lead to the reasonable expectation that a BERT-based classification model would outperform previously used methodologies. Since 2018, several augmentations occurred, with Facebook publishing the RoBERTa language model in 2019,<sup>10</sup> surpassing previously set records. The RoBERTa is a late, robust, unsupervised pre-trained language model that can be used in the context of supervised tasks with outstanding results.<sup>11</sup>

Undoubtedly, the abundance of clinical information is locked in clinical narratives. Documentation of EHRs is now developing into standard practice. For instance, surgeons spend significant time documenting and reading, amongst other tasks, narrative descriptions of operative reports and findings.<sup>12</sup> Developing tools to facilitate clinical review of

these unstructured data can derive clinically meaningful insights for advanced epithelial ovarian cancer (EOC), a heterogeneous disease. Compared to standard approaches, they can potentiate condensation of results from several tasks and optimize analysis time. One aspiration could be the prediction of no residual disease (R0 resection) following cytoreductive surgery for EOC. Such task of confirming macroscopic clearance remains subjective,<sup>13</sup> to the point that photographic 'mapping' has been recommended that allows for an assessment of the surgical effort at primary surgery or provides a baseline for determining the effect of neo-adjuvant chemotherapy at delayed surgery.<sup>14</sup> As a result, most of the quantitative intraoperative assessment tools have mainly focused on their predictive value for suboptimal surgery.<sup>15</sup> To improve modern care, the application of NLP tools could be useful to determine whether processing of unstructured full-text documents improves the ability to forecast outcomes in clinical conditions with significant heterogeneity such as EOC.

In this work, we utilized the pre-trained RoBERTa-base language model to predict whether residual disease persists in EOC patients following their cytoreductive surgery. We hypothesized that operative notes contain valuable information associated with surgical outcomes. We aimed to develop an NLP methodology that would address the objectiveness of R0 resection through information hidden in unstructured operative notes.

## Methods

Electronic Health Records (EHRs) were queried to identify women with advanced EOC who underwent cytoreductive surgery at St James's University Hospital, Leeds, from January 2014 to December 2019. The modern EHR dataset included the following clinical features: diagnosis codes (ICD-10 codes), procedure codes (OPCS-4 codes), age at diagnosis, grade, stage, and operative notes with findings. An internally developed advanced EOC clinical database was integrated with the EHR system<sup>16</sup> to provide the availability of discrete and engineered data. Institutional research ethics board approval was obtained through the Leeds Teaching Hospitals Trust (MO20/133 163/18.06.20), and informed written consent was obtained. The study was added to the UMIN/CTR Trial Registry (UMIN000049480). Treatment was pre-operatively planned at the weekly central gynaecological oncology multidisciplinary team (MDT) meeting prior to patient review. The cohort details, hospital setting, indications for surgery, and surgical procedures have been described in our previous studies.<sup>13,17</sup> Comprehensive visual assessment of all the areas

of the abdomen and pelvis was routinely performed, and no visible residual disease was documented as R0 resection. The analysis took place in three steps: Firstly, words and combinations of words were analyzed based on their frequency and the case they concerned. Following the initial descriptive text analysis, the RoBERTa classifier was employed to predict case outcomes based on operative notes. Lastly, an XGBoost classification model was tasked with predicting the same outcome, this time using tabular discrete data, but also the probabilities that were derived from the RoBERTa classifier of the second step. A flowchart of our approach is shown in Figure 1.

### Textual Descriptive Analysis

For the analysis of the text, word frequencies were calculated, and tables were created using the most common words and n-grams. N-grams are continuous word sequences of words, as they could be found in the text. The length of the n-grams can be as small as one, meaning one word, or as large as the entirety of the text. N-grams are important because they carry contextual information more than simple words do. To find the n-grams that best discriminated between the two cases, we performed an analysis based on the Term Frequency – Inverse Document Frequency (TF-IDF). The TF-IDF is a metric used to quantify n-gram importance in a particular document.<sup>18</sup> The score, as implied by its name, is a function of the number of times the n-gram appears in the document adjusted for the number of times it appears in the rest of the documents, as shown also in equation (1).

$$tfidf_{i,j} = tf_{i,j} * \log \frac{N}{df_i} \quad (1)$$

where

- (i)  $tf_{i,j}$  is the number of occurrences of n-gram  $i$  in document  $j$
- (ii)  $df_i$  is the number of documents containing  $i$
- (iii)  $N$  is the total number of documents

For each of the two possible outcomes (R0 resection vs non-R0 resection), we compiled a document consisting of the concatenation of all the individual notes that concerned this outcome. The words inside the documents were reduced to their lemmas, to make the analysis more representative of the real n-gram frequency without accounting for word conjugation. The two resulting documents were inputted into Sklearn's TfidfVectorizer, which was tasked with assigning scores per n-gram, per document. A high TF-IDF score for an n-gram in a document signifies n-gram importance to this document. The TF-IDF n-gram scores for the documents reporting non-R0 resection were then subtracted from the TF-IDF scores for the documents reporting R0 resection. This way, the higher the absolute difference in scores, the higher the ability of the n-gram to discriminate between the two cases. Positive difference scores show that the n-gram belongs to R0 resection case notes, while negative scores show the opposite.

### Natural Language Classification With RoBERTa

We utilized the pre-trained RoBERTa-base language model to extract information from the unstructured surgical notes through transfer learning. Transfer learning is the process of re-training part of a pre-trained model on specific data to fine-tune its performance for a specific task. The initial training

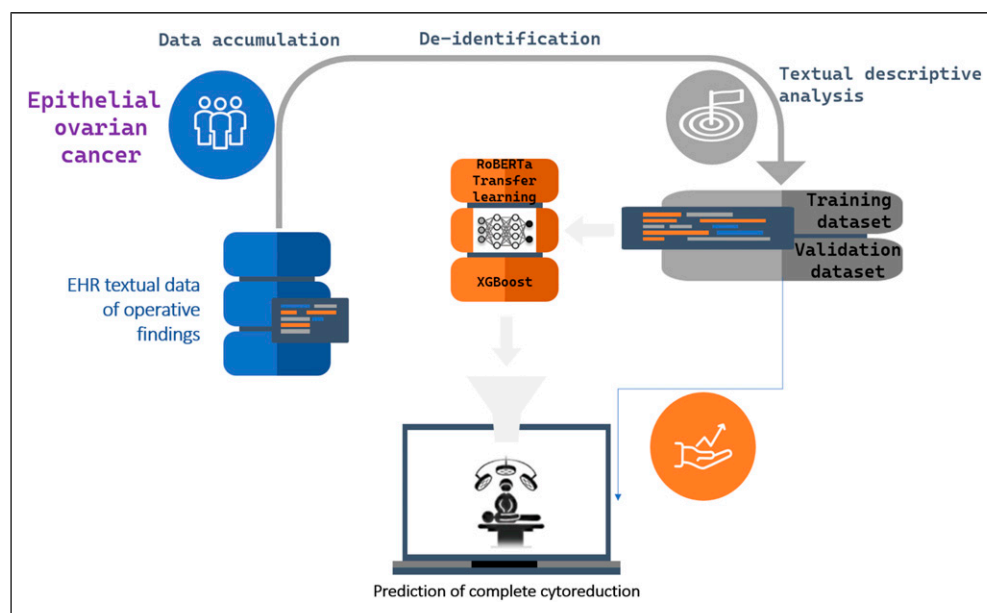


Figure 1. Components and the flow of the machine learning pipeline applied in our case.

often uses vast datasets that hold most of the information relevant to the task at hand. Re-training allows for finer details to be captured by the model. The main advantage of transfer learning is that the resulting model can reach high performance without needing to use large amounts of data. The RoBERTa-base language model is pre-trained on a large corpus of English data using the BERT-base architecture and has 125 million parameters.<sup>10</sup> The surgical data was used to train and test the model at a ratio of 4:1. The model was trained for 40 epochs. Discrimination was measured using the most common performance metrics for classification tasks, namely, with:

- (i) Accuracy =  $\frac{TP + TN}{TP + TN + FP + FN}$
- (ii) Precision =  $\frac{TP}{TP + FP}$
- (iii) Recall =  $\frac{TP}{TP + FN}$
- (iv) F1-score =  $\frac{2 * Precision * Recall}{Precision + Recall}$
- (v) Area under the receiver operating characteristic curve (AUC)
- (vi) Area under the precision-recall curve (AUPRC)

Understanding how the RoBERTa reached the conclusions is essential in evaluating its performance. For such an extremely complicated model, tracing the individual impact of the text tokens on the final prediction is a task requiring advanced compositional methods. In this effort, we employed the transformers-interpret Python library<sup>19</sup> to explain and visualize the factors that contributed to the model's prediction accuracy. In turn, the library employs the Captum model of interpretability and understanding library.<sup>20</sup> Using integrated gradients, the library evaluates the contribution of each input feature to the model output of the model. The net result is an attribution score for each token; that is positive when the token contributes towards class prediction and negative in the reverse scenario.

As a final step, we employed a surrogate model in order to augment the explainability effort. In this context, a surrogate model denotes a simpler model than the powerful original one (in our case RoBERTa), whose outputs are interpretable. The surrogate model trains on the outputs of the original and, through its interpretable coefficients, offers a way to access the decision process of the complex, original model. The surrogate model used was a simple logistic regression. The dependent variable was the RoBERTa predictions in the form of binary integer values, created by setting a threshold of .5 on the original probabilities, while the independent variables were the TF-IDF sentence vectors created through the method of TF-IDF vectorization. In this way, the aim was 2-fold: The surrogate model would serve both as an explainer and as a conceptual link between the RoBERTa outputs and the TF-IDF scores created in the descriptive analysis.

### XGBoost Classification Model

Subsequently, we trained an XGBoost model<sup>21</sup> to predict R0 resection using a combination of structured and unstructured

data sources. The independent variables included the Aletti surgical complexity score (SCS), the size of the largest bulk of the disease in centimeters, the age of the patient, the Pre-Surgery CA125, the Intraoperative Mapping of ovarian cancer (IMO) score, the operative time in minutes, the estimated blood loss (EBL), the pre-treatment CA125, the tumour grade encoded as a binary variable, the Peritoneal Carcinomatosis Index (PCI), the timing of surgery (encoded as a binary variable where primary debulking equaled 0 and interval debulking surgery equaled 1), the ANAFI score and the probabilities that the RoBERTa classifier outputted when solely tasked to predict R0 resection (real number in the interval of 0 to 1). The PCI and IMO scores were calculated at the beginning of surgery to describe the intraoperative location of the disease.<sup>22,23</sup> The Aletti SCS was assigned to describe the surgical effort.<sup>24</sup> The ANAFI score is an AI-derived novel intraoperative score that assigns specific weights to the EOC dissemination patterns (ANatomic Fingerprint).<sup>25</sup> It appears to be more predictive of R0 resection than the entire PCI and IMO scores whilst it retains its prognostic power. Most of these discrete and engineered data predictors have been interrogated in our previous studies.<sup>13,17,25-27</sup>

The hyperparameters of the XGBoost model were selected by using an exhaustive grid hyperparameter search. The grid search also implemented cross-validation. The hyperparameter grid is shown in Table 3. The feature importance was determined using the Shapley additive explanations (SHAP) framework to interpret the model's predictions based on the Shapley values.<sup>28</sup>

## Results

Using the ICD-10 code for EOC, we identified 555 cases of EOC cytoreduction performed by eight surgeons between January 2014 and December 2019. This cohort has been previously described.<sup>13,17</sup> Some basic descriptive statistics are shown in Table 1. The rate of complete cytoreduction was 65.4%.

### Textual Descriptive Analysis

Discrete word clouds weighted by n-gram TF-IDF score difference (Table 2) between R0 and non-R0 resection were identified (Figure 2).

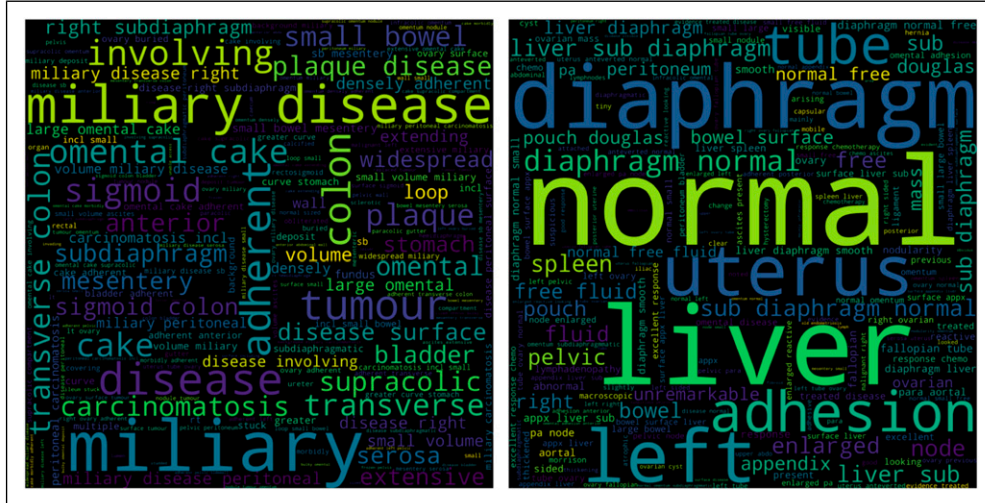
The words 'normal' and 'miliary' best discriminated between both groups. For non-R0 resection prediction, these included n-grams related to the EOC dissemination, such as 'omental cake.' The appearance of the cancer was best described by the predictive n-gram 'miliary disease.' The average word count was  $320 \pm 98$  vs  $292 \pm 105$  in case notes with non-R0 vs R0 resection, respectively. The average stop word count was  $11.22 \pm 6.27$  vs  $9.86 \pm 5.89$  in case notes with non-R0 vs R0 resection, respectively.

**Table 1.** Cohort Statistics.

Variable	Overall (n = 555)	Training Set (n = 444)	Testing Set (n = 111)	P-Value (Training)	Zero Residual (n = 363)	Non-Zero Residual (n = 192)	P-Value (R0 vs Non-R0)
Grade_cat	502 (90.45)	403 (90.77)	99 (89.19)	.745	332 (91.46)	170 (88.54)	.337
IDS/PDS	385 (69.37)	306 (68.92)	79 (71.17)	.730	247 (68.04)	138 (71.88)	.404
Surgical complexity score (SCS)	3.77 ± 2.07	3.8 ± 2.04	3.66 ± 2.2	.544	4.13 ± 2.27	3.08 ± 1.4	<.001
Size largest bulk of disease (cm)	8.83 ± 5.57	8.71 ± 5.62	9.3 ± 5.37	.306	8.31 ± 5.67	9.79 ± 5.25	.002
Age	63.57 ± 11.23	63.88 ± 10.96	62.32 ± 12.25	.221	62.45 ± 11.65	65.69 ± 10.1	.001
Pre-surgery CA125	412.99 ± 1180.36	406.44 ± 1229.0	439.45 ± 963.76	.762	399.52 ± 1297.67	438.6 ± 919.64	.682
Intraoperative mapping of ovarian cancer (IMO)	4.9 ± 1.95	4.83 ± 1.91	5.2 ± 2.11	.096	4.36 ± 1.88	5.93 ± 1.65	<.001
Time procedure (min)	168.82 ± 75.13	169.5 ± 72.33	166.08 ± 85.71	.699	172.77 ± 80.26	161.35 ± 63.84	.068
EBL	521.84 ± 386.84	523.06 ± 400.16	516.95 ± 329.82	.868	512.1 ± 417.74	540.26 ± 320.62	.377
Pre-treatment CA125	1525.33 ± 2719.94	1421.98 ± 2573.98	1938.7 ± 3218.95	.118	1499.46 ± 2911.22	1574.24 ± 2322.0	.742
PCI	7.3 ± 4.39	7.16 ± 4.26	7.89 ± 4.85	.145	6.52 ± 4.3	8.79 ± 4.17	<.001
ANAFI	5.02 ± 5.45	4.72 ± 5.27	6.23 ± 5.98	.016	2.85 ± 4.4	9.13 ± 4.86	<.001

**Table 2.** Top 10 n-Grams With the Highest TF-IDF Difference Scores per Case Outcome. The Two Top-Level Header Columns Indicate the Case According to Which the Score Difference was Sorted. R0 n-Grams had High Positive Score Difference While Non-R0 n-Grams had High Negative Score Difference.

Word	TF-IDF Score for Word in R0 Document	TF-IDF Score for Word in Non-R0 Document	TF-IDF Score Difference
<b>R0</b>			
Normal	.261	.154	.106
Liver	.151	.068	.082
Diaphragm	.141	.071	.070
Left	.213	.164	.048
Uterus	.213	.168	.044
Adhesion	.108	.064	.044
Tube	.094	.052	.042
Diaphragm normal	.046	.006	.039
Sub-diaphragm normal	.045	.006	.038
Liver sub	.046	.007	.038
<b>Non-R0</b>			
Miliary	.026	.170	-.143
Miliary disease	.018	.127	-.109
Disease	.269	.346	-.077
Tumour	.108	.178	-.069
Adherent	.170	.237	-.066
Colon	.058	.116	-.058
Involving	.012	.059	-.046
Omental cake	.031	.073	-.042
Cake	.032	.073	-.040
Sigmoid	.090	.131	-.040



**Figure 2.** N-gram word clouds for findings notes where residual disease is non-zero (left) and zero (right).

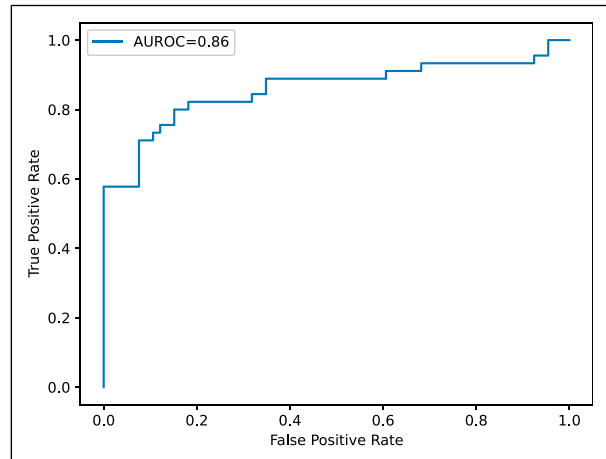
### Natural Language Classification With RoBERTa

The model reached high evaluation metrics (area under ROC .86; area under precision-recall curve .87, precision, recall, and F1 score of .77 and accuracy of .81 [Figures 3–5]), surpassing even specialized BERT and DistilBERT models tested (BioBERT<sup>29</sup>: R .6, P 0.84, F1 .7, ACC .79, AUROC .84, AUPRC .84, ClinicalBERT<sup>30</sup>: R .68, P 0.72, F1 .7, ACC .76, AUROC .82, AUPRC .82 and BioClinicalBERT<sup>31</sup>: R .64, P 0.76, F1 .69, ACC .77, AUROC .81, AUPRC .79). The true positives, true negatives, false positives and false negatives were 35, 56, 10 and 10, respectively.

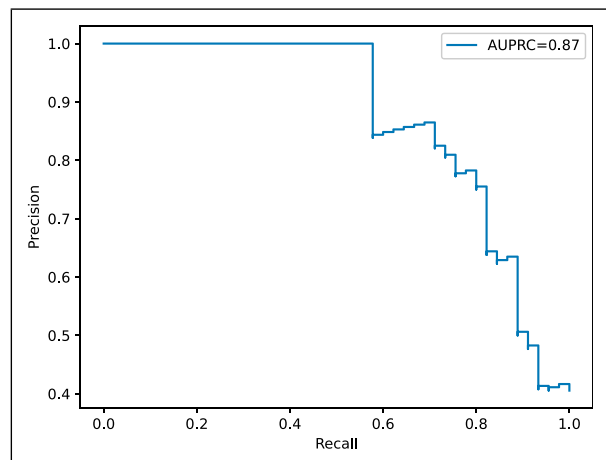
The explanations of the model’s predictions were visualized in a highlighted text plot, where negatively contributing tokens are coloured red, and positively contributing tokens are coloured green, whereas the colour intensity is translated to an attribution score strength (Figure 6).

As shown in the figure, it is easy to discern the fact that there is a correlation between word contribution to the prediction and TF-IDF score difference. This makes sense, as n-grams with high TF-IDF score difference tend to discriminate better between the two cases. However, it is equally important to see that not all words that have high prediction contribution score appear as entries in Table 2. That is due to the fact that since RoBERTa is able to capture contextual meaning spanning several words that could also be non-sequential, it is possible that local information that was not apparent through simple TF-IDF analysis was now deemed as important to the prediction.

The results of the surrogate logistic regression model employed further reinforce the results acquired from the RoBERTa model. Specifically, n-grams that reached high TF-IDF difference scores (Table 2) appeared as top coefficients in the logistic regression, in either direction (Figure 7).



**Figure 3.** Receiver operating characteristic curve and area under the curve for the RoBERTa classifier.



**Figure 4.** Precision-recall curve and area under the curve for the RoBERTa classifier.

## XGBoost Classification Model

The XGBoost model that employed both discrete features and the probabilities from the RoBERTa classifier was then trained on the same training data set as the RoBERTa. The grid hyperparameter search resulted in 180 model evaluations with the best combination of hyperparameters shown in Table 3 alongside the search spaces. For the XGBoost model, while the precision, recall, F1 score, and accuracy remained static, a marginal performance improvement was demonstrated as shown by the AUPRC and AUROC reaching .91. The RoBERTa probabilities, when used as a prediction feature, performed significantly better than not only discrete features but also engineered features such as the ANAFI score (Figure 8).

## Discussion

In this proof-of-principle study, we demonstrated the capability of the RoBERTa classifier to extract and process

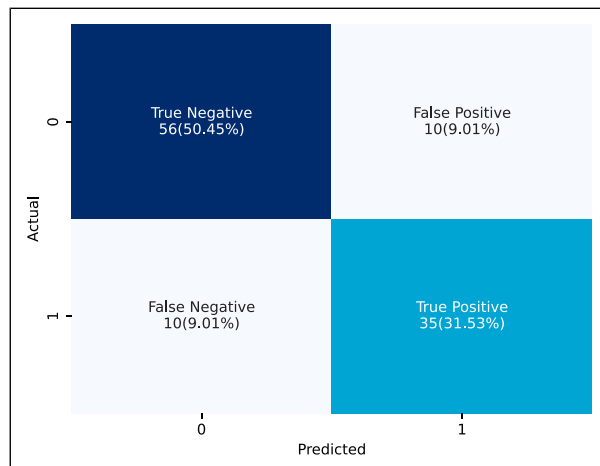


Figure 5. Confusion matrix for the RoBERTa classifier.

information from unstructured operative note formats that can enable important clinical tasks, such as R0 resection prediction following EOC surgical cytoreduction. We showcased how EHRs can be a helpful data source for supporting surgeons' activities by automated data coding for quality assessment while reducing the burden of chart review. As an estimated 70% of clinicians report EHR-related, specialty-specific burnout,<sup>32</sup> this information may guide healthcare organizations on how to remediate burnout amongst their staff. Equally, we surmise this effort can help establish interoperability standards of surgical narration to ensure objectivity when it comes to reporting residual disease. Working with EHR data is relatively challenging due to data heterogeneity. Being able to quickly retrieve important information stored in surgical narratives carries the potential to improve understanding of patient journeys and identify subgroups of patients for research purposes. For those reasons, the design and application of a system that could offer the NLP AI-derived insights directly to the surgeon in real-time would be extremely beneficial. The system could offer objective feedback on written notes. A study on the effects of such a system should be investigated.

The driving motivation behind this effort was to explore the potential of using the RoBERTa algorithm in the EOC domain. This transformer architecture has been recently used to extract adverse drug events from biomedical text to monitor drug safety.<sup>33</sup> Barber et al initially developed an NLP-augmented algorithm that improved the ability to predict postoperative complications and hospital readmissions among women with EOC undergoing surgical cytoreduction.<sup>34</sup> They compiled discrete data with different types of NLP features from unstructured clinical notes and sequentially employed machine learning to build new sets of features. Herein, we purely used a novel NLP tool that recognizes the specific local textual context, thus enabling a recommendation concerning the prediction of residual disease. Pre-processing steps contributed

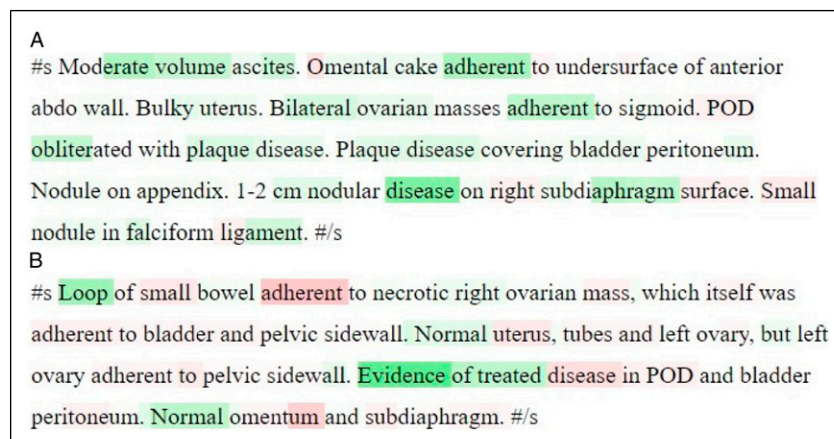
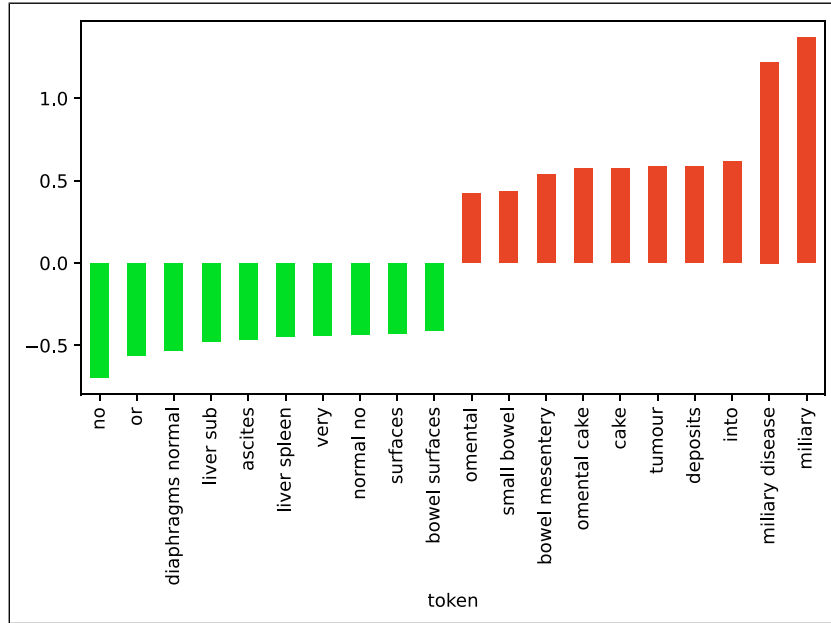


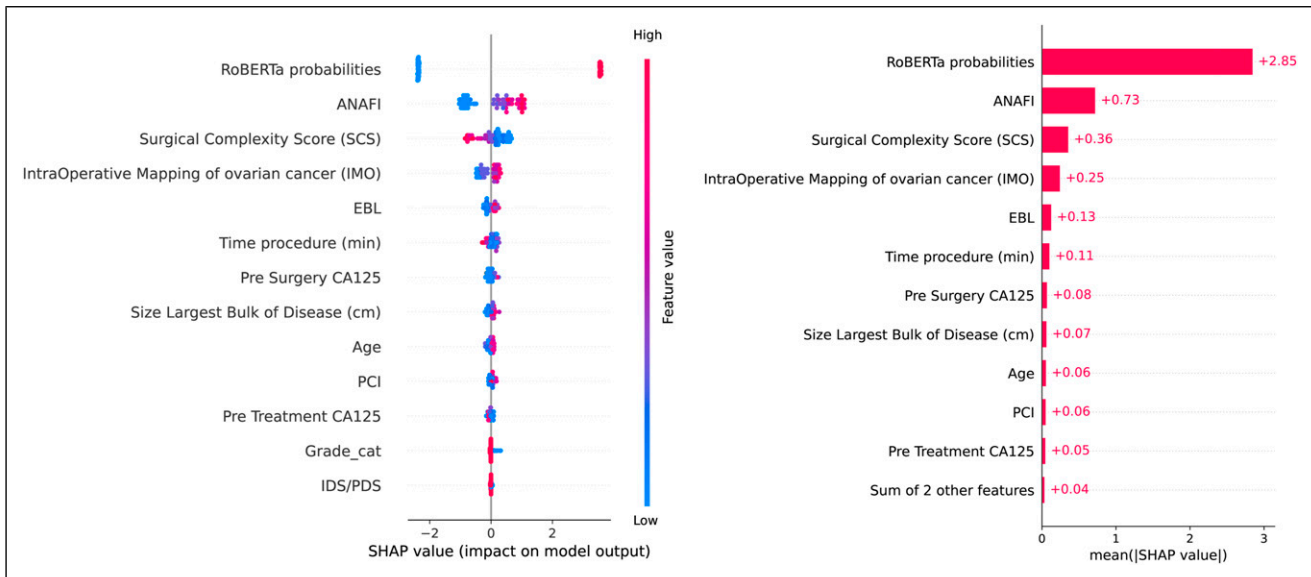
Figure 6. Explainability of the RoBERTa inference on textual data. The green highlighting indicates the section of text that contributed positively to the classification of the note as belonging to the assigned class, while the red highlighting indicates the opposite. Examples are text instances correctly classified as describing cases where (A) residual disease persisted and (B) no residual disease persisted after surgery.



**Figure 7.** Top 10 n-grams with the lowest (green) and highest (red) coefficients of the logistic regression model. The negative sign denotes non-existence of residual disease and vice versa.

**Table 3.** Hyperparameter Search Space of Grid Search and Chosen Parameters.

Hyperparameter	Search Space	Chosen Hyperparameter Value
max_depth	[4, 6, 8, 10]	4
n_estimators	[500, 800, 1200]	500
learning_rate	[.01, .03, .08]	.01
colsample_bytree	None	0.5



**Figure 8.** Explainability plots for the XGBoost classification model. The beeswarm feature impact plot (left) visualizes the relationship between direction of model prediction and value of feature. The bar feature impact plot (right) shows the absolute impact of each feature on model prediction.



to the rather high AUROC of .86, which shows how surgeons tend to capture more of the predictive information in their words. This ‘hunch’ critically layered upon situational awareness, and human factors have been addressed in our previous study.<sup>13</sup> The model specificity was higher than its sensitivity, which is critical, should this be used as a cancer screening tool for quality control. Reports of surgical findings are less restrictive in vocabulary than other EHRs, but their efficiency at scale has never been previously examined. They do not usually contain highly complex sentence structures, so they are not incorrectly abstracted as a result. By avoiding the model to make assumptions, this advantage would potentially explain the high-performance accuracy.

More importantly, we demonstrated a distinct pattern of word differential expression between R0 resection and non-R0 resection operative notes from 555 surgical events. While survival is the ultimate treatment outcome, prediction of residual disease is a key issue in the advanced EOC trajectory. This disease quantification can valuably complement our previous work using AI to predict EOC-specific surgical outcomes<sup>13,17,25–27</sup> and validate the paradigm shift towards complete clearance to improve the survival outcomes of these patients.<sup>13,35</sup> The use of language in medicine is often underestimated not that all Gynaecologic Oncology Surgeons speak the same language.<sup>36</sup> Historically, the quantification of both peri-operative disease burden and post-operative residual disease in advanced EOC was subject to significant intra- and inter-observer variability, particularly in the case of miliary peritoneal disease. While addressing the need to improve standardization and reproducibility of surgical outcomes, we made some interesting observations. Despite several words or n-grams being commonly shared between examined surgical outcomes, several descriptive words were found to be predictive of residual disease. For instance, the words ‘stuck’ and ‘adherent’ tend to describe a more complex and morbid surgery; dissemination leading to residual disease was best described by ‘(small volume) miliary disease’ or ‘miliary in all peritoneal surfaces’. ‘Excellent response to chemo’ was clearly an obvious indication to achieve R0 resection. Not surprisingly, words demonstrating hepatobiliary involvement were referring to those patients who had had macroscopically complete resection of all visible tumours.<sup>37</sup> Notably, the ‘completion of cytoreduction’ (CC) scoring system was developed to evaluate the extent of resection for peritoneal malignancies.<sup>22</sup> We clearly showed that the word ‘miliary’, if quantified, rather refers to CC1 (residual disease nodules up to 2.5 mm in size) of the perhaps outdated Sugarbaker classification (PCI).<sup>22</sup> We provide valid language evidence that the CC score is more likely to give a convincing and reproducible description of residual disease in EOC. In addition, the subtle performance superiority of textual data when compared with discrete surgical data can be also invoked. Going forward, data integration between structured and unstructured formats can

promote innovative thinking to perfect the prediction of surgical outcomes, offer important indications for the treatment of patients and contribute to policies and clinical guidelines with the goal of reducing the future risk of unnecessary morbidity and mortality.<sup>38</sup>

The challenges of Machine Intelligence in healthcare have been consistently addressed.<sup>39</sup> Transfer learning requires a close collaboration between clinicians and computer scientists. Until that happens, the inherent resilience in these tools will delay their widest adaptation. We anticipate the portability of our RoBERTa algorithm across similar practice settings by conducting original studies albeit we acknowledge the heterogeneous nature of the clinical language. Historically, linguistic models are evaluated by perplexity, that is, the probability of predicting the word in its context. Our study used retrospective data from a single institution. On that note, our data size was small-to-moderate, which entertains the general wisdom that the fewer data feeds the model, the higher the perplexity can get. Our important observations were made in a tertiary referral center; hence, they might not be generally applicable. We reiterate our strong preference for explainable NLP methods,<sup>40</sup> which has been showcased in this work. Understanding the features that drive a model prediction can potentially support decision-making in the healthcare domain. As NLP is moving to deep learning, it is becoming increasingly challenging for these complex non-linear data transformations to satisfy transparency.<sup>41</sup>

The latest hype from the technological advancements in large language models has been embraced with some cautious excitement. Undoubtedly, AI-based chatbots engage in a capacity to understand multiple languages and possess knowledge of various topics. They can generate fabricated information in healthcare settings word by word.<sup>42</sup> In ovarian cancer research, most efforts focus on addressing the disease heterogeneity.<sup>43</sup> It is likely that this heterogeneity contains ‘special grammars’ that cannot be distilled from simply vast amounts of pre-trained textual data resources. Our work highlights the need for a bespoke, proprietary ovarian cancer-specific natural language that can pay attention to detail and learn beyond human knowledge.

## Conclusion

We applied a sui generis approach to extract the information from the abundant textual surgical data through the use of an NLP model utilizing transfer learning and demonstrated how such tasks can be effectively modeled for the classification of prediction important surgical tasks, such as R0 resection following advanced EOC surgical cytoreduction. State-of-art NLP applications in biomedical texts can improve modern EOC care.

## Acknowledgments

We wish to thank all the staff caring for our ovarian cancer patients at LTHT.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Ethical Statement

### Ethical Approval

Institutional research ethics board approval was obtained through the Health Research Authority (IRAS application 2823960) and the Leeds Teaching Hospitals Trust (MO20/133163/18.06.20).

### Statement of Human and Animal Rights

The author(s) state(s) that this research was conducted in accordance with the Helsinki Declaration as revised in 2008.

### Informed Consent

Informed consent was obtained for the study.

## ORCID iDs

Alexandros Laios  <https://orcid.org/0000-0002-4870-7393>  
 Marios Evangelos Mamalis  <https://orcid.org/0009-0000-2680-0442>  
 Diederick De Jong  <https://orcid.org/0000-0003-0081-674X>

## References

- Uppal S, Al-Niaimi A, Rice LW, et al. Preoperative hypoalbuminemia is an independent predictor of poor perioperative outcomes in women undergoing open surgery for gynecologic malignancies. *Gynecol Oncol.* 2013;131(2):416–422. doi:10.1016/j.ygyno.2013.08.011 <https://www.sciencedirect.com/science/article/pii/S0090825813010962>
- Barber EL, Rutstein SE, Miller WC, Gehrig PA. A preoperative personalized risk assessment calculator for elderly ovarian cancer patients undergoing primary cytoreductive surgery. *Gynecol Oncol.* 2015;139(3):401–406. doi:10.1016/j.ygyno.2015.09.080 <https://www.sciencedirect.com/science/article/pii/S0090825815301414>
- Modi S, Feldman SS. The value of electronic health records since the health information technology for economic and clinical health act: Systematic review. *JMIR Med Inform.* 2022;10(9):e37283. doi:10.2196/37283
- Martin-Sanchez F, Verspoor K. Big data in medicine is driving big changes. *Yearb Med Inform.* 2014;9(1):14–20.
- Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. *JMIR Med Inform.* 2020;8(3):Article e17984. doi:10.2196/17984 <http://www.ncbi.nlm.nih.gov/pubmed/32229465>
- Zhu Y, Mahale A, Peters K, et al. Using natural language processing on free-text clinical notes to identify patients with long-term covid effects. In Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. Number 46 in BCB '22, New York, NY, USA: Association for Computing Machinery, p. 9. doi:10.1145/3535508.3545555
- Seol HY, Sohn S, Liu H, et al. Early identification of childhood asthma: The role of informatics in an era of electronic health records. *Front Pediatr.* 2019;7:113.
- Devlin J, Chang MW, Lee K, et al. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171–4186. *Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv; 2018.
- Barber EL, Garg R, Persenaire C, Simon M. Natural language processing with machine learning to predict outcomes after ovarian cancer surgery. *Gynecol Oncol.* 2021;160(1):182–186. doi:10.1016/j.ygyno.2020.10.004
- Liu Y, Ott M, Goyal N, et al. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach (cite arxiv:1907.11692). *Roberta: A Robustly Optimized Bert Pretraining Approach.* ArXiv; 2019.
- Wang Alex, Singh Amanpreet, Michael Julian, Felix Hill, Omer Levy, Samuel Bowman. 2018. *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.* In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: A perspective on the tension between structure and flexible documentation. *J Am Med Inf Assoc.* 2011;18(2):181–186. doi:10.1136/jamia.2010.007237 <https://academic.oup.com/jamia/article-pdf/18/2/181/6035310/18-2-181.pdf>
- Laios A, Kalampokis E, Johnson R, et al. Factors predicting surgical effort using explainable artificial intelligence in advanced stage epithelial ovarian cancer. *Cancers.* 2022;14(14):3447.
- Jones M, Mohamed F. Photography transillumination techniques: Multicystic peritoneal mesothelioma. *J Biocommun.* 2020;44(1):e3.
- Hosoya S, Ueda K, Odajima S, et al. Scoring systems of peritoneal dissemination for the prediction of operative completeness in advanced ovarian cancer. *Anticancer Res.* 2022;42(1):115–124.
- Newsham AC, Johnston C, Hall G, et al. Development of an advanced database for clinical trials integrated with an electronic patient record system. *Comput Biol Med.* 2011;41(8):575–586. doi:10.1016/j.combiomed.2011.04.014

17. Laios A, Kalampokis E, Johnson R, et al. Explainable artificial intelligence for prediction of complete surgical cytoreduction in advanced-stage epithelial ovarian cancer. *J Personalized Med*. 2022;12(4):607.
18. Tang H, Ng JHK. Googling for a diagnosis—use of google as a diagnostic aid: Internet based study. *BMJ*. 2006;333(7579):1143-1145. doi:10.1136/bmj.39003.640567.ae
19. Pierser C, Paaß, G., Giesselbach, S. (2023). Pre-trained Language Models. In: Foundation Models for Natural Language Processing. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, Cham. [https://doi.org/10.1007/978-3-031-23190-2\\_2](https://doi.org/10.1007/978-3-031-23190-2_2). *Transformers Interpret*. Springer. <https://github.com/cdpierser/transformers-interpret> (2021).
20. Kokhlikyan Narine, MiglaniVivek, Martin Miguel, Wang Edward, AlsallakhBilal, Reynolds Jonathan, Melnikov Alexander, Kliushkina Natalia, Araya Carlos, Yan Siqi, Reblitz-Richardson Orion. (2020). Captum: A unified and generic model interpretability library for PyTorch.
21. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, New York, NY, USA: ACM, pp. 785-794. doi:10.1145/2939672.2939785 <http://doi.acm.org/10.1145/2939672.2939785>
22. Jacquet P, Sugarbaker PH. Clinical research methodologies in diagnosis and staging of patients with peritoneal carcinomatosis. In: *Cancer Treatment and Research*. Berlin: Springer; 1996: 359-374. doi:10.1007/978-1-4613-1247-5\_23
23. Sehoul J, Könsgen D, Mustea A, et al. ["IMO"—intraoperative mapping of ovarian cancer]. *Zentralblatt für Gynäkologie*. 2003; 125(3/4):129-135. doi:10.1055/s-2003-41864
24. Aletti GD, Dowdy SC, Podratz KC, Cliby WA. Relationship among surgical complexity, short-term morbidity, and overall survival in primary surgery for advanced ovarian cancer. *Am J Obstet Gynecol*. 2007;197(6):e1-e7. DOI:10.1016/j.ajog.2007.10.495
25. Laios A, Kalampokis E, Johnson R, et al. Development of a novel intra-operative score to record diseases' anatomic fingerprints (ANAFI score) for the prediction of complete cytoreduction in advanced-stage ovarian cancer by using machine learning and explainable artificial intelligence. *Cancers*. 2023; 15(3):966. doi:10.3390/cancers15030966
26. Laios A, Katsenou A, Tan YS, et al. Feature selection is critical for 2-year prognosis in advanced stage high grade serous ovarian cancer by using machine learning. *Cancer Control*. 2021;28: 10732748211044678.
27. Laios A, De Freitas DLD, Saalmink G, et al. Stratification of length of stay prediction following surgical cytoreduction in advanced high-grade serous ovarian cancer patients using artificial intelligence; the leads I-AI-OS score. *Curr Oncol*. 2022; 29(12):9088-9104. doi:10.3390/curroncol29120711
28. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable ai for trees. *Nat Mach Intell*. 2020;2(1):56-67.
29. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36(4): 1234-1240. doi:10.1093/bioinformatics/btz682. Feb 15.
30. Wang G., Liu X., Ying Z., et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nat Med* 2023. <https://doi.org/10.1038/s41591-023-02552-9>
31. Alsentzer E, Murphy JR, Boag W et al. *Publicly Available Clinical Bert Embeddings*, 2019. 1904.03323.arXiv:1904.03323 [cs.CL]. <https://doi.org/10.48550/arXiv.1904.03323>
32. Gardner RL, Cooper E, Haskell J, et al. Physician stress and burnout: The impact of health information technology. *J Am Med Inf Assoc*. 2019;26(2):106-114.
33. Jain H, Raj N, Mishra S. A sui generis QA approach using RoBERTa for adverse drug event identification. *BMC Bioinf*. 2021;22(Suppl 11):330.
34. Barber EL, Garg R, Persenaire C, Simon M. Natural language processing with machine learning to predict outcomes after ovarian cancer surgery. *Gynecol Oncol*. 2021;160(1):182-186.
35. De Jong D, Otify M, Chen I, et al. Survival and chemosensitivity in advanced high grade serous epithelial ovarian cancer patients with and without a BRCA germline mutation: More evidence for shifting the paradigm towards complete surgical cytoreduction. *Medicina*. 2022;58(11):1611. doi:10.3390/medicina58111611
36. Brennan DJ, Moran BJ. Time to evolve terminology from “debulking” to cytoreductive surgery (CRS) in ovarian cancer. *Ann Surg Oncol*. 2021;28(11):5805-5807. doi:10.1245/s10434-021-10490-4
37. Di Donato V, Giannini A, D’Oria O, et al. Hepatobiliary disease resection in patients with advanced epithelial ovarian cancer: Prognostic role and optimal cytoreduction. *Ann Surg Oncol*. 2021;28(1):222-230.
38. Benedetti Panici P, Giannini A, Fischetti M, Lecce F, Di Donato V. Lymphadenectomy in ovarian cancer: Is it still justified? *Curr Oncol Rep*. 2020;22(3):22.
39. NiOBla B. *Machine Intelligence in Healthcare: Nih*; 2019. <https://ncats.nih.gov/expertise/machine-intelligence#workshop>
40. Chen L, Gu Y, Ji X, et al. Clinical trial cohort selection based on multi-level rule-based natural language processing system. *J Am Med Inf Assoc*. 2019;26(11):1218-1226.
41. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform*. 2018;22(5):1589-1604. doi:10.1109/jbhi.2017.2767063
42. Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*. 2023;11(6):887. doi:10.3390/healthcare11060887
43. Hu Z, Artibani M, Alsaadi A, et al. The repertoire of serous ovarian cancer non-genetic heterogeneity revealed by single-cell sequencing of normal fallopian tube epithelial cells. *Cancer Cell*. 2020;37(2):226-242. doi:10.1016/j.ccell.2020.01.003