To predict or not to predict? The dilemma of predicting the risk of suboptimal cytoreduction in ovarian cancer

S. Kang & S.-Y. Park*
Center for Uterine Cancer, National Cancer Center, Goyang, Republic of Korea

Although maximal cytoreduction is the cornerstone of current treatment for patients with advanced ovarian cancer, optimal cytoreduction is not always achievable in the clinic. Therefore, using clinical characteristics, diagnostic imaging, serum biomarkers or laparoscopic findings, many studies have attempted to find models for predicting surgical resectability. However, most of these prediction models showed limited effectiveness and have not been properly validated. To establish a reliable prediction model, several requirements should be met. First, the goal of surgical cytoreduction should be adequately defined. Second, the desired accuracy for making the model clinically useful should be defined. Third, the model should test all relevant predictors, including clinical, radiological and biochemical predictors, and be developed using a large dataset that provides a sufficient number of events. Fourth, any prediction model should be validated with a relevant external dataset. Lastly, the prediction model should be able to aid decision making and, thereby, improve the outcome of patients. Therefore, randomized clinical trials with decision making based on prediction models are urgently required.

Key words: ovarian cancer, surgery, optimal cytoreduction, residual tumor, prediction, neoadjuvant chemotherapy

introduction

Ovarian cancer is the leading cause of gynecologic cancer death in Western countries [1]. At least two-thirds of women who are diagnosed with ovarian cancer have advanced disease at the time of diagnosis. Currently, the standard treatment for these patients is maximal cytoreductive surgery followed by platinum-based systemic chemotherapy [2]. After Griffiths reported that cytoreduction to <1.5 cm in size is an important prognostic factor [3], gynecologic oncologists have struggled to achieve a higher rate of optimal cytoreduction. The cut-off for determining optimal cytoreduction has been changed repeatedly [4–7] and a residual tumor size of <1 cm is now generally accepted as the definition of optimal surgical cytoreduction [7].

However, despite maximal surgical effort, optimal cytoreduction is not always possible in patients with advanced disease [8, 9]. Therefore, many studies have attempted to develop a model to predict the possibility of optimal cytoreduction, using various clinical characteristics, diagnostic imaging, laparoscopic findings and biomarkers such as CA125 [10]. However, none of these models has successfully elicited a consensus in the clinical community and many gynecologic oncologists remain skeptical about the feasibility of preoperative prediction of optimal cytoreduction [11]. Why have so many studies failed to convince gynecologic oncologists to use their prediction tools in the clinic? Why are gynecologic oncologists reluctant to use these proposed prediction tools in their practice?

reference standard: how should the term ‘optimal’ be defined?

First, the concept of optimal cytoreduction has not been clarified, which introduces significant confusion in developing a prediction model of optimal cytoreduction. In the early studies that sought to predict the surgical resectability of ovarian cancer, optimal cytoreduction was defined as a residual tumor of <2.0 cm in diameter [12, 13]. However, recently, most studies define optimal cytoreduction as a residual tumor of <1.0 cm [14, 15]. Moreover, many experts now claim that not 1.0 cm, but complete resection, i.e. the absence of a gross residual tumor, should be the goal of the surgical management of ovarian cancer [6, 16, 17]. In a recent exploratory analysis of three prospective randomized trials (AGO-OVAR 3, 5 and 7), du Bois et al. reported that the overall survival of advanced ovarian cancer patients significantly differed between those without a gross residual tumor and patients with a residual tumor of <1.0 cm [17]. On the other hand, the authors observed that the impact of debulking the residual tumor to <1.0 cm was very small. The patients with residual tumors of >1.0 cm showed a hazard ratio of 1.2 compared with patients with residual tumors of 1–10 mm. Thus, considering the impact on survival outcome, it would be more appropriate to define optimal cytoreduction as the complete absence of a gross residual tumor. The cut-off that defines the ‘optimal’ size of the residual tumor in prediction models may not only determine the design and
performance of a prediction model, but also influence the acceptance of the model by clinicians. Thus, to develop a clinically useful prediction model, the definition of ‘optimal’ cytoreduction should be developed and clearly addressed.

**accuracy: how should a prediction tool be evaluated?**

The accuracy of a prediction model is the most important issue determining its usefulness in clinical decision making. A prediction model is unlikely to be useful unless its predictive power is at least as high as that of the doctor who would use it. However, unsurprisingly, it is confusing and often frustrating to evaluate and compare the many suggested models for predicting the surgical resectability of ovarian cancer. Many studies have provided various quantitative indicators of prediction performance, including accuracy, sensitivity, specificity, predictive values, likelihood ratios (LRs) and area under the receiver operating characteristic (ROC) curve. Among these indicators, accuracy and predictive values are the most commonly used indicators in most studies. However, because performance indicators such as accuracy or predictive values are dependent on the prevalence of events, and because the rate of optimal cytoreduction varies among the study cohorts, it is evident that these indicators are not adequate for evaluating the performance of models predicting surgical resectability [18]. However, indicators such as sensitivity or specificity are not highly intuitive and clinicians may therefore find it difficult to understand the clinical usefulness of a model based on these indicators. Thus, many studies frequently transform these indicators into LRs or diagnostic odds ratios (19, 20). A positive LR can be calculated using the following formula: sensitivity / (1 – specificity). A positive LR is defined as the probability of positive test results among patients with events divided by the probability of positive test results among patients without events. Usually, a positive LR >5.0 indicates that the test is useful for predicting positive outcomes [21, 22]. Conversely, a negative LR can be calculated using the formula: (1 – sensitivity) / specificity. Similarly, a negative LR <0.2 indicates that the test can be applied for excluding negative outcome [21, 22]. In Table 1, the sensitivity and specificity of various studies investigating the performance of computed tomography (CT) imaging in predicting surgical resectability were transformed into positive and negative LRs [12, 13, 15, 23–28]. Six of the nine studies resulted in a positive LR >5.0, suggesting that many studies have reported CT imaging to be useful for identifying patients with a high risk of suboptimal cytoreduction. Two studies showed high positive LRs and low negative LRs, suggesting that the overall performance of these models may be strong. Thus, any prediction model for surgical resectability should be evaluated by performance indicators resistant to variability in event prevalence.

**accuracy: how accurate is accurate enough?**

It should be noted that many prognostic studies in the cancer-research field based on a multivariate modeling approach yield

| Table 1. Proposed diagnostic imaging-based prediction model and validation results |
|-----------------|----------|-------|-----------|-----------|
| N               | Sensitivity (95%) | Specificity (95%) | Positive LR | Negative LR |
| Nelson et al. [12] | 42        | 92.3  | 79.3      | 4.5       | 0.097     |
| Meyer et al. [13] | 28        | 58    | 100       | –         | 0.42      |
| Bristow et al. [15] | 41    | 100   | 85        | 6.7       | 0         |
| Byrom et al. [28] | 51        | 88    | 92        | 11.0      | 0.13      |
| Qayyum et al. [26] | 137       | 76    | 99        | 76.0      | 0.24      |
| Dowdy et al. [27] | 87        | 52    | 90        | 5.2       | 0.53      |
| Astell et al. [25] | 65        | 79    | 75        | 3.2       | 0.28      |
| Jüng et al. [23] | 77        | 43    | 80        | 2.2       | 0.71      |
| Ferrandina et al. [24] | 195 | –    | –        | 10.3      | –         |

LR, likelihood ratio. Positive LR = sensitivity / (1 – specificity); negative LR = (1 – sensitivity) / specificity.
accuracy: issue of overfitting

Many of the studies on surgical resectability had small numbers of subjects and thus had a very limited number of events. When the size of the dataset is limited, the number of candidate predictors that can be reliably tested in a multivariate analysis is also limited. A well-known rule of thumb is the 1-in-10 rule, i.e. 1 candidate predictor can be studied for every 10 events [35, 36]. For example, for 30 patients who underwent suboptimal cytoreduction in a study of 100 patients, only 3 candidate predictors could reliably be tested. However, this rule has been frequently violated in studies on prediction models for surgical resectability, especially in studies testing too many image-based predictors. If the rule is violated, the number of candidate predictors is too large for the dataset and overfitting is likely to occur. This issue subsequently results in a situation where a model that performed well in the original dataset poorly reproduces in an independent dataset. A recent study analyzed medical data and CT images of 65 patients, and the event, i.e. suboptimal cytoreduction, was observed only in 14 out of 65 patients in the study [25]. The authors selected two predictors: diaphragmatic disease and mesenteric involvement. They achieved an accuracy of 77% by combining these two predictors into the model. However, if we look at the dataset, we can easily see that the authors would achieve the same accuracy if they selected only one of the two predictors while following 1 : 10 rule. The model reproduced poorly in external datasets, which may have been a consequence of overfitting. Therefore, it is crucial to reduce the risk of overfitting in the development stage of a risk model. In this context, we have some concerns about risk assessment using laparoscopy. In general, it is likely that an overestimation of surgical outcome may be introduced when a surgeon is informed of the test results before surgery. Thus, many risk assessments using laparoscopy may be constrained by concerns about overestimation of their performance.

generalizability: the importance of external validation

Many clinicians could be reluctant to adopt a prediction model derived from one specific population or institution, because of a generalizability issue. Thus, adequate validation of a model in an independent external dataset is very important to minimize the risk of overoptimism. Axell et al. sought to investigate the diagnostic performance of their own prediction model in two independent cohorts [25]. In addition, they tested the diagnostic performances of two other previously developed models based on CT predictors in their cohort. The authors found that the performance of the tested models reproduced poorly in the independent dataset. Performance discrepancies among different institutions have also been observed in many studies testing the predictive performance of serum CA125 [37]. In a previous meta-analysis reported by Kang et al., the predictive ability of CA125 >500 IU/ml to predict optimal cytoreduction varied among institutions. Moreover, two studies from the same institution but at a different time and under different surgical policies resulted in different predictive performances [14, 38]. However, there is a model that showed a similar performance in an external validation. Based on laparoscopy findings, Fagotti et al. derived a prediction scoring system with an overall accuracy of 75% [34]. Unlike the validation study of various imaging models, the accuracy of the laparoscopy-based model was similarly reproduced in an independent patient set [39], suggesting that developing a generalizable model for resectability is not an unachievable goal.

what factors should be considered?

To convince clinicians of the credibility of a prediction model, it is important that all clinically relevant patient data have been tested for inclusion in the model. Factors used to predict surgical resectability might include clinical characteristics of the patients, the extent and size of the tumor in diagnostic imaging, serum levels of biomarkers, or the results of direct visual examination by laparoscopy. First, it may be wise to include and test the age of patients and performance status as predictors, although the relationship between these parameters and optimal cytoreduction is debatable [24, 28, 40]. There has been little evidence suggesting that certain histologic types are associated with the difficulty of cytoreductive surgery. However, one study suggested that some minor histologic types were more likely to be completely removed than a serous type [41, 42]. The most important predictors may be various indicators of the anatomic spread, extent and size of the tumor. Usually, these parameters are investigated as a form of imaging data. Among the many predictors that have been suggested, a few commonly suggested predictors are listed here: diaphragmatic disease [15, 23–25], a large volume of ascites [15, 23, 27], disease involving the mesentery [15, 24, 25] and diffuse peritoneal thickening or implants [15, 27]. Although some of these factors could merely be correlated with other predictors rather than true predictors, each of these predictors should be thoroughly tested when attempting to establish a prediction model. Another tool for assessing the risk of suboptimal cytoreduction may be direct visual examination via laparoscopy. Because the accuracy of diagnostic imaging is not perfect, some researchers have attempted to develop a scoring system based on laparoscopic findings. Fagotti et al. designed a scoring system consisting of eight parameters [34]. Although it can be conceded that laparoscopy may be the most accurate prediction tool, it is questionable whether a clinical decision-making process based on such an invasive procedure would result in more benefits than costs, including the risk of port-site metastasis [43]. Lastly, serum biomarkers, such as CA125, can be used to predict resectability. Predicting resectability based on CA125 was first suggested by Chi et al. [14] and led many researchers to investigate the hypothesis in their own patient populations [27, 38, 44–57]. Kang et al. summarized these results in a recent meta-analysis and concluded that serum CA125 has a sensitivity of 69% and specificity of 63% at a cut-off of 500 IU/ml [37]. Although a serum CA125 level >500 IU/ml is a strong risk factor of suboptimal cytoreduction (odds ratio = 3.7), the biomarker alone was not useful for the prediction of optimal cytoreduction. Therefore, the ideal method for establishing a risk model may be to test all the above candidate parameters.
increased survival gain [62]. More importantly, the randomized increased rate of optimal cytoreduction did not translate into a prognostic factor in advanced ovarian cancer. However, optimal cytoreduction [61], which is the most important neoadjuvant chemotherapy significantly increased the rate of assumption was based on a number of observations that models for predicting optimal cytoreduction, many researchers treatment is there to improve her outcome? While investigating a patient is determined to be unresectable, what alternative therefore, what kind of medical decision can be made? If a treatment has a survival outcome that is similar to a standard treatment but likely to offer reduced morbidity, decreased blood loss and shorter hospital stays, why should we make an effort to classify a group of patients who are suitable for the standard treatment rather than using this alternative treatment liberally? In this context, predicting surgical resectability may not influence the decision of either advocates or opponents of neoadjuvant chemotherapy. Therefore, not only do we currently lack a reliable tool for assessing the surgical resectability of ovarian cancer, but we also do not have a rationale for tailoring a patient’s treatment according to the predicted resectability. Although it is unclear whether current prediction models are useful in clinical decision making in terms of survival benefit or operative morbidity, it may help to anticipate the complexity of surgery and the need to perform extensive surgery to achieve optimal residual disease [38]. Therefore, whether a prediction model of resectability may aid the decision to transfer patients to a high-volume cancer center [68–71] and thereby enhance the survival of ovarian cancer patients should be evaluated further.

conclusions

Maximal cytoreductive surgery is the cornerstone of treatment for advanced ovarian cancer. However, it is also true that some patients have unfavorable risk factors indicating difficult optimal surgery. Although we do not have any evidence that clinical decision making based on the prediction of surgical resectability can improve patient outcome, a relevant risk
model is required to assess patient risk, compare study populations and design clinical trials for triage. For this purpose, first, a consensus should be reached regarding the goal of surgical cytoreduction. Second, the optimal accuracy of the model should be defined. Third, the model should test all relevant predictors and be developed using a sufficiently large dataset providing enough events. Fourth, for external validation, any prediction model should provide reproducibility in a relevant external dataset outside the institution where the model was initially developed. Fifth, the prediction model should be tested as to whether it can aid decision making and improve patient outcome, including quality of life or morbidity. Any appropriate hypothesis should be suggested and robustly discussed within the community of gynecologic oncologists. In addition, these hypotheses should be tested in well-designed, randomized controlled trials. To achieve these goals, communication and cooperation among international oncologic study groups are necessary.

acknowledgements

This work is supported by the National Cancer Center, Korea (grant no. 0910260-3).

disclosures

The authors declare that there is no conflict of interest.

references

6. Chi DS, Eisenhauer EL, Lang J et al. What is the optimal goal of primary cytoreductive surgery for bulky stage IIC epithelial ovarian carcinoma (EOC)? Gynecol Oncol 2006; 103: 559–564.


