Premiere Publications from The Triological Society

Read all three of our prestigious publications, each offering high-quality content to keep you informed with the latest developments in the field.

**Laryngoscope**
*Founded in 1896*

Editor-in-Chief: Michael G. Stewart, MD, MPH
The leading source for information in head and neck disorders.
[Laryngoscope.com](http://Laryngoscope.com)

**Laryngoscope Investigative Otolaryngology**

Editor-in-Chief: D. Bradley Welling, MD, PhD, FACS
Rapid dissemination of the science and practice of otolaryngology-head and neck surgery.
[InvestigativeOto.com](http://InvestigativeOto.com)

**ENTtoday**

Editor-in-Chief: Alexander Chiu, MD
Must-have timely information that Otolaryngologist-head and neck surgeons can use in daily practice.
[Enttoday.org](http://Enttoday.org)

WILEY
Meta-Analysis of Induction Chemotherapy as a Selection Marker for Chemoradiation in the Head and Neck

Kimberley L. Kiong, MMed (ORL) ©; Nurun Nisa de Souza, MD, MPH; Rehena Sultana, M.Sc; N. Gopalakrishna Iyer, PhD, FRCSEd

Objective: Many trials incorporate induction chemotherapy (IC) in selecting for organ preservation in head and neck squamous cell carcinomas (HNSCC). However, few studies examine IC response in predicting for chemoradiation therapy (CRT) response. This meta-analysis aims to determine the predictive accuracy of IC for subsequent response to CRT and overall survival (OS).

Data sources: Medline, EMBASE, Cochrane register.

Methods: A systematic search identified studies from database inception to October 2016 that used IC prior to CRT as definitive treatment for advanced HNSCC. The sensitivities and specificities of IC response predicting for complete CRT response were calculated, and the results were pooled in a summary receiver operating curve. One-, 2-, and 5-year OS data were extracted.

Results: Seven studies (n = 423 patients) were analyzed for response and six (n = 439) for OS. Pooled median sensitivity and specificity of IC response predicting CRT response were 0.95 (95% confidence interval [CI]: 0.72–0.98) and 0.43 (95% CI: 0.00–0.61), respectively. Patients were more likely to respond to CRT given previous response to IC (positive likelihood ratio = 1.6; 95% CI: 1.21–2.11) and less likely to respond to CRT if they failed to respond to IC (negative likelihood ratio = 0.16; 95% CI: 0.07–0.38). At 2 years, good response to IC was a statistically significant prognostic marker with a risk ratio of 1.35 (95% CI: 1.12–1.64).

Conclusion: Our data suggests that patients with poor IC response will have poorer response to CRT and should be directed to other modalities. In contrast, good IC response does not guarantee a favorable outcome to CRT; however, because these patients are likely to have better prognoses, they should be offered salvage therapies of curative intent despite treatment failure.

Key Words: Chemotherapy, adjuvant, chemoradiotherapy, head and neck neoplasms.
Level of Evidence: NA.

Laryngoscope, 128:1594–1601, 2018

INTRODUCTION

Organ preservation is one of the key objectives in non-surgical management of head and neck squamous cell cancers (HNSCC) in a bid to preserve vital functions such as speech and swallowing.1 The landmark Veteran's Affairs (VA) larynx preservation study led the paradigm shift of using chemoradiation therapy to achieve this endpoint, although extending this concept across all HNSCC has not been straightforward.2 Certainly, in cancers arising from subsites such as the oral cavity, randomized data suggests that concurrent chemoradiation therapy (CRT) appears to be inferior to primary surgery.3

Apart from organ preservation, results of the VA study also drove a shift toward widespread use of neoadjuvant or induction chemotherapy (IC). The latter was purported to have important implications on treatment selection, planning, and prognostication. In most solid tumors, the value of IC is to downsize tumors to facilitate resection and organ preservation,4,5 with a secondary intent of reducing the burden of micrometastasis even before primary treatment.6,7 For HNSCC, induction protocols from the VA study were quickly superseded by CRT as the standard of care for organ preservation.8 More recently, protocols incorporate the sequential addition of IC to CRT, moving from single-agent regimes to multiple cytotoxic drugs. Certainly, in this context the addition of IC appears to have an effect on decreasing the incidence of distant metastasis.9–11 Furthermore, many protocols use IC as a form of chemoselection to identify patients who would likely respond to CRT. In fact, a number of randomized trials have used response to IC in their decision tree: Good responders go on to receive CRT, whereas poor responders are offered surgery upfront.12,13 This assumes that a good response to
IC indicates a likely good outcome to CRT and vice versa.

Although this approach is intuitive and has been demonstrated in some individual studies, the notion that response to IC predicts ultimate response to CRT and overall prognosis\(^{14,15}\) has yet to be conclusively proven. Proponents cite the lack of good biomarkers for CRT selection and contend that the best biomarker for response is response. In contrast, others argue that using chemoselection merely serves to identify a cohort that has poor prognosis regardless of treatment given.\(^{16}\)

Systematic reviews testing this hypothesis have been lacking, and individual studies have been underpowered to answer this question specifically. There is anecdotal evidence to challenge these assumptions, even in the landmark trials. For example, in the radiation therapy oncology group (RTOG) 91-11 trial, 24 patients were reported to have poor response to IC and were recommended total laryngectomy but did not proceed with surgery.\(^{16}\) Instead, 11 of these patients received additional CRT, and all achieved complete response. Similarly, in a study by Lim et al. looking specifically at a group who showed poor response to IC, patients who went on to RT/CRT (for various reasons) showed an overall response rate of 61.5%.\(^{17}\) It is important to note the majority of data supporting the argument that response to IC predicts radiosensitivity arises from an era in which definitive RT rather than CRT was used. Moreover, these were largely based on single-agent IC regimes, which have been superseded by multi-agent protocols. As such, using response to IC as part of a chemoselection biomarker in the current setting remains unproven.

Therefore, in this systematic review and meta-analysis, our primary objective was to investigate whether successful treatment with multi-agent IC, as defined by complete and partial response, predicts a similar response to CRT in patients with locally advanced HNSCC. Our secondary objective was to determine whether IC response predicts overall survival in this cohort.

METHODS

Literature Search and Quality Assessment

Literature Search Strategy. A literature search was performed on MEDLINE, EMBASE, and the Cochrane Central Register of Controlled Trials. The search period was from database inception until October 2016. Key terms searched included “head and neck cancer,” “induction chemotherapy,” and “sequential chemoradiation.” Both indexed and free text terms were included in the search strategy (Appendix 1, available online). Only articles published in English were searched. Reference lists of included articles and relevant systematic reviews were also searched to identify potentially relevant articles.

Study Selection. Studies were included if they recruited patients with American Joint Committee on Cancer stage III and IV HNSCC without distant metastases, had a treatment protocol that employed IC followed by definitive concomitant CRT, and reported response to both IC and CRT. Response was reported as complete response (CR), partial response (PR), stable disease (SD), or progressive disease (PD) based on World Health Organization (WHO) criteria or Response Evaluation Criteria in Solid Tumors (RECIST) criteria.\(^{18}\) Studies in which subjects had recurrent HNSCC or were previously treated for HNSCC were excluded. Studies that focused on the use of experimental treatment or biological/immune therapies such as cetuximab were also excluded. Two reviewers (K.L.K. and N.N.Đ.) independently assessed articles for eligibility by title, abstract, and full text. Disagreements were resolved through discussions.

Data Extraction. Data from the eligible studies were extracted using a standard data extraction template. The primary endpoint was overall response to CRT, and the secondary endpoint was overall survival (OS). The data collected included the study design, tumor characteristics, IC and CRT regimens, and response to IC and CRT and survival outcomes. Two reviewers (K.L.K. and K.S.) verified all extracted data, and any disagreements were resolved through discussion. When an agreement could not be reached, a third author (N.N.Đ.) was conferred with until a final agreement could be achieved.

When studies only reported OS through Kaplan–Meier survival curves, Engauge Digitizer version 4.1 was used to extract quantitative data from the reported curves. We did not encounter missing data that required us to contact the authors for clarification.

Quality Assessment of Bias. Two reviewers (K.L.K. and N.N.Đ.) independently assessed for methodological quality using A Cochrane Risk of Bias Assessment Tool for Non-Randomized Studies of Interventions.\(^{19}\) For each study, seven domains were assessed. Each was given a judgment of low, moderate, serious, or critical risk of bias before an overall risk of bias judgment was determined.

Statistical Analysis and Quantitative Synthesis

We defined a CR or PR as a good (positive) response to IC, and SD or PD as poor (negative) response to IC. Only CR to CRT was deemed a positive result, and anything less than a CR to CRT was deemed as a negative result. Based on these definitions, a \(2 \times 2\) table based on responses to IC and CRT was formed. The numbers of true positives, true negatives, false positives, and false negatives were obtained from that table.

For the secondary outcome, we extracted OS at 12, 24, and 60 months. We then developed a table for the number of patients surviving and total number of patients in CR/PR and SD/PD separately for each time point.

Data Synthesis. The summary estimate of sensitivity, specificity, positive likelihood ratio (PLR), and negative likelihood ratio (NLR) with corresponding 95% confidence interval (CI) was obtained. We conducted meta-analyses of correlated pairs of sensitivity and specificity using both hierarchical summary receiver operating curve (HSROC)\(^{20,21}\) and bivariate model.\(^{22}\) The size of the point depicts the precision of the estimate scaled according to the sample sizes, with height relating to the precision of sensitivity and width relating to the precision of the specificity estimate. The scatter of point estimates showed that both sensitivity and specificity presented similar variability. The diagnostic odds ratio (DOR), which presents a summary of the diagnostic accuracy of the test, describes the ratio of the odds of a positive IC test result for a patient who ultimately achieves CR to CRT as opposed to a patient who achieves less than CR with CRT. Because we expected heterogeneity in a meta-analysis of diagnostic accuracy, we used the random effects model.

Sensitivity and specificity are unique to the test was being used. Because of threshold effect, sensitivity and specificity are expected to be very heterogeneous. In combination with the correlation between sensitivity and specificity, this made testing
for heterogeneity using the I^2 statistic problematic. Hence for the primary outcome, we explored heterogeneity by visual inspection of paired forest plot of sensitivity and specificity, and of HSROC plot.

For the secondary outcome, we combined extracted results using a random effects model and expressed pooled results in terms of risk ratios (RRs) with corresponding 95% CI. We also conducted subgroup analysis based on 12, 24, and 60 months as sensitivity analysis. I^2 statistics was used to evaluate the heterogeneity of overall test accuracy between the studies. I^2 equal to 0% meant no heterogeneity; 0% < I^2 < 40% meant not important heterogeneity; 30% < I^2 < 60% meant moderate heterogeneity; 50% < I^2 < 90% meant substantial heterogeneity; and I^2 > 75% meant considerable heterogeneity.

**Software.** A paired forest plot of sensitivity and specificity and a forest plot of overall survival were made using Review Manager version 5.3 (Cochrane Collaboration, Oxford, U.K.). HSROC, bivariate SROC, and accuracy of the test curve were plotted using MetaDas macro in SAS version 9.3 (SAS Institute, Cary, NC).

**RESULTS**

**Search Results and Study Characteristics**

A flow diagram of the study selection process is provided in Figure 1. The literature search identified 6,356 citations. After removal of duplicates, 5,416 citations remained. These were screened resulting in 19 studies that met the inclusion criteria. One study that reported response to IC/CRT in terms of metabolic response in PET scans and not by the traditional imaging criteria was excluded. Kim et al. reported outcomes including progression free survival but did not report OS, and was excluded. Studies by Ampil et al., Chang et al., and Kim et al. did not report the individual responses to CRT stratified by CR/PR versus SD/PD responses to IC and were also excluded. One other study looked specifically at patients who achieved SD/PD to IC and excluded patients with CR/PR response to IC. Hence, 13 studies that recruited a total of 862 patients were deemed eligible and included in this meta-analysis (Supporting Table SI). Of these, seven studies were analyzed for the primary outcome of response rates and six studies for the secondary outcome of OS. The IC protocols used in these studies were multi-agent regimes consisting of between two to three chemotherapeutic agents, most commonly taxanes, platinum-based drugs, and 5-flourouracil (5FU). Some studies included agents such as methotrexate, and the number of cycles given varied between one to six cycles.

**Study Quality Assessment**

All except three studies were were rated as having moderate to serious risk of bias. In most of the studies, the study quality was compromised due to lack of adequate adjustments for confounding in their analyses, inherent in their nonrandomized nature. In addition, one study considered surgery in patients with incomplete response to IC or CRT in their analysis and did not give the exact number of patients who underwent surgery. Details of the study quality assessment are described in Figures 2A and 2B.

**Response to Induction Chemotherapy as Predictor of Response to Chemoradiation Therapy**

Seven studies with 423 patients had available data on specific response rates to IC and CRT. One study focused on neck disease only, and one study reported the breakdown of response rates in the primary and neck sites separately, which were analyzed separately.

The median sensitivity and specificity for good response to IC predicting for complete response to CRT were 0.95 (maximum: 0.98; minimum: 0.72) and 0.43 (maximum: 0.61; minimum: 0.00), respectively. Specificity showed a greater uncertainty than sensitivity, indicated by CI width, because the number of noncases in most of the studies is generally lower than the number of cases (Fig. 3).

Figure 4 presents the ROC scatter plot displaying the results of sensitivity and specificity for individual studies in the ROC space, where each of the eight data sets is plotted as a single sensitivity-specificity point with corresponding 95% confidence region. The black circle depicts point estimate of the summary sensitivity and specificity. The prediction region was large, indicating considerable between-study heterogeneity. No clear pattern was observed with regard to the location of studies above or below the ROC curve, depending on their size.

The PLR was 1.60 (95% CI: 1.21–2.11) and indicated that a good responder to IC (“test positive”) is 1.6 times more likely to respond fully to CRT than a poor responder (“test negative”). The NLR was 0.16 (95% CI: 0.07–0.38), with resulting DOR of 9.77 (95% CI: 3.51–27.17). The NLR represents the likelihood that a poor responder to IC will respond to CRT compared to a poor responder to IC failing to respond to CRT. The DOR, another measure of the accuracy of response to IC as a

---

Fig. 1. Flow diagram of study selection.
test for response to CRT, represents the ratio of the odds of CRT response in good IC responders relative to the odds of CRT response in poor IC responders. None of the covariates, for example, site and country, were found to be statistically significant as a possible source of heterogeneity.

Fig. 2. (A) Risk of bias graph. Review authors' judgements about each risk of bias item presented as percentages across all included studies. (B) Risk of bias summary. Review authors' judgements about each risk of bias item for each included study.
Response to IC as a Predictor of Survival

Six studies (439 patients) analyzed the 12 months OS of patients stratified according to good/poor response to IC. Five studies reported 24 months OS, whereas three had data for 60 months OS. At 12 and 24 months, subjects who achieved a good response to IC were more likely to survive at 12 months (RR = 1.20, 95% CI: 1.06–1.35), 24 months (RR = 1.35, 95% CI: 1.12–1.64), and 60 months (RR = 1.88, 95% CI: 0.75–4.71). Survival at 60 months was not significant, and the 95% CI was wide. This was likely due to the small number of studies that reported OS at 60 months. Combining these data, good response to IC was found to be a significant prognostic marker for better overall survival with a RR of 1.26 (95% CI: 1.12–1.40) (Fig. 5). We attempted to analyze other survival indices, but there were too few studies that reported such data.

DISCUSSION

The hypothesis of our review is that IC serves as a form of in vivo assay to identify patients who would likely respond to CRT. Also, successful response to IC may result in altered tumor kinetics, with subsequent better response to CRT. The rationale for multi-agent IC is to target a heterogeneous tumor population at different points of vulnerability. For example, cisplatin and taxanes/taxol target cells at different cell cycle checkpoints, whereas 5FU functions to inhibit metabolic function during the synthesis phase of the cell cycle. The timing of these agents, together with fractionated radiotherapy, is theoretically designed to synchronize tumor kinetics across the entire treatment protocol such that most if not all tumor cells enter the vulnerable state during the course of treatment. These assumptions, however, do not take into account other effects of cytotoxics, such as activation of signaling pathways and immune modulation. All the above has led to the incorporation of IC into trial designs. Moreover, current clinical practice is also underpinned by this notion.

In this study, the focus was to determine the accuracy with which multi-agent IC regimes predict for ultimate response to CRT. Based on the studies included, our data demonstrates that good response to IC has a high sensitivity of 95% to identify responders, but a relatively low specificity of 43% in predicting for eventual complete response to CRT. In clinical terms, a NLR of 0.16 indicates that patients with SD/PD to IC have a 30% to 45% higher probability of not achieving CR to CRT, and hence supports protocols where these patients should be directed to primary surgery. In contrast, a PLR of 1.60, which does suggest a higher probability that those with CR/PR to IC will respond to CRT, is not particularly high and confers less than 15% higher probability of CR to CRT. These are estimates of probability, and are likely to be accurate to within 10% of the calculated answer for all pretest probabilities between 10% and 90% with an average error of only 4%. We obtained a summary DOR of 9.77 (95% CI: 3.51–27.17), another measure of test performance that combines the strengths of sensitivity and specificity, independent of prevalence. The value and CIs of more than 1 indicate that response to IC as a test is discriminatory for response to CRT. However, because no other “tests” for response to CRT are available, there is no means of comparison of this value.

Fig. 3. Sensitivities and specificities of good response to induction chemotherapy as a test for complete response to chemoradiation therapy. [Color figure can be viewed at www.laryngoscope.com.]

Fig. 4. Summary receiver operator characteristics curve of good response to induction therapy as a test for complete response to chemoradiation therapy.
For a multi-disciplinary tumor board making decisions, these results suggest that poor responders to induction chemotherapy are less likely to benefit from CRT and should hence direct patients to other therapeutic options. The value of using a good IC response to predict good response to CRT is more limited. However, the caveat in this analysis is the definition of response, which was that CR/PR was deemed to be good for IC, whereas only a complete response was deemed ideal for CRT.

Survival outcomes based on response to IC consistently showed that a good response to IC (CR/PR) predicts for better overall survival at 1, 2, and 5 years, with a trend toward higher RRs over time. This has been supported by a number of individual studies but not by previous systematic reviews.\textsuperscript{35,37–40} The prognostic value herein supports current practice guidelines because response to IC is still a useful biomarker of prognosis, and those who fail CRT may still go on to salvage surgery to achieve good outcomes.

Despite the large number of trials that evaluate IC in HNSCC, we could only analyze 13 studies for several reasons. A significant proportion of studies direct poor IC responders to primary surgery as part of their treatment protocol, an assumption that is supported by our findings. However, these studies were excluded because the objective was to correlate the biological assumption that CRT response replicates IC response. The value of other agents such as cetuximab was not evaluated because there were limited studies, and many similar agents are not used routinely. Furthermore, many of the larger trials did not include a breakdown of numbers in terms of response rates, nor did they stratify outcomes according to IC response. This resulted in a smaller number of studies and patients being included, reflecting only a small percentage of the trials that perform IC followed by CRT, which is one of the major limitations of our meta-analysis. Also, the studies lack individual patient specific data, and none of the studies reported both our outcomes of interest.

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>CR/PR Events</th>
<th>SD/PD Events</th>
<th>Total Events</th>
<th>Weight</th>
<th>Risk Ratio IV, Random, 95% CI</th>
<th>Risk Ratio IV, Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - year survival</td>
<td>26 31 5 8 3.8%</td>
<td>70 73 26 31 25.6%</td>
<td>19 26 6 10 3.6%</td>
<td>24 26 5 10 3.0%</td>
<td>144 157 20 26 18.1%</td>
<td>345 94 57.4%</td>
</tr>
<tr>
<td>Total events</td>
<td>307 67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Heterogeneity: Tau^2 = 0.00; Chi^2 = 2.44, df = 5 (P = 0.79); I^2 = 0% | Test for overall effect: Z = 2.98 (P = 0.003)

| 2 - year survival | 25 31 4 8 2.4% | 64 73 21 31 14.1% | 8 26 3 10 1.0% | 23 32 3 9 1.4% | 123 157 15 26 9.1% | 319 84 27.9% | 2.98 [1.00, 8.48] | 1.29 [1.00, 1.67] | 1.03 [0.34, 3.11] | 2.16 [0.63, 7.57] | 1.36 [0.97, 1.91] | 1.35 [1.12, 1.64] |
| Total events | 243 46 | | | | | | 3.11 [1.26, 7.60] |
| Heterogeneity: Tau^2 = 0.00; Chi^2 = 1.51, df = 4 (P = 0.82); I^2 = 0% | Test for overall effect: Z = 1.35 (P = 0.18)

| 5 - year survival | 48 73 21 31 11.5% | 17 26 1 10 0.4% | 104 157 7 26 2.9% | 2.46 [1.29, 4.68] | 1.88 [0.75, 4.71] | |
| Total events | 169 29 | | | | | | 1.35 [1.12, 1.60] |
| Heterogeneity: Tau^2 = 0.46; Chi^2 = 9.90, df = 2 (P = 0.007); I^2 = 80% | Test for overall effect: Z = 1.35 (P = 0.18)

| Total (95% CI) | 920 245 100.0% | 1.26 [1.12, 1.40] | |
| Total events | 719 142 | | | | | | 3.98 (P < 0.0001) | Test for subgroup differences: Chi^2 = 1.94, df = 2 (P = 0.38), I^2 = 0%

**Fig. 5.** Meta-analysis of risk ratio for overall survival according to response to induction therapy. [Color figure can be viewed at www.laryngoscope.com.]
Another limitation arose from the heterogeneity of the studies. The studies included in our meta-analysis used a combination of two to three drugs for IC (Supporting Table SI). Because there is no established regime for IC, the available studies varied widely on type of and number of cycles (2–6) of IC. Our original intent was to perform subgroup analyses on the effects of the IC regimen itself on response and outcome. However, given the small number of studies and the fact that they were considerably heterogeneous, we were unable to perform any useful meta-analysis. Also, the definitions used to evaluate response rate (e.g. WHO criteria, RECIST criteria) and evaluation intervals varied, contributing to the heterogeneity. However, this is a reflection of the existing literature that we currently can use to draw conclusions and the need for better trials to establish an answer.

The practice of using IC as chemoselection is set to continue because there is no other biomarker predictive of response to CRT. Numerous previous studies have attempted to identify potential biomarkers, but results have proven disappointing. Initial results from analyzing patients from the VA trial suggested immunohistochemistry of proteins such as p53 and Bcl-xL as promising targets, but these were not borne out in larger studies. However, given the current availability of high-throughput, multi-dimensional datasets, this question should be revisited because it is certainly an important unmet clinical need.

CONCLUSION

In conclusion, the objective of this study was to establish the value of response to IC to subsequent CRT response and survival. To our knowledge, this is the only meta-analysis that specifically asks this question. Our findings suggest that patients with poor IC response will have poor response to CRT and should therefore be directed to other treatment modalities, such as (but not limited to) primary surgery. In contrast, a good response to induction therapy does not guarantee a favorable outcome to CRT; however, given that these patients have better prognosis, they can be offered salvage therapies of curative intent even after treatment failure. Future studies are needed to evaluate the role of newer classes of therapeutics, such as biologic agents, which may further be able to influence our armamentarium. There continues to be a need to identify and test multi-dimensional biomarker panels to predict CRT response from the outset.

Acknowledgment

We would like to thank Serene AK Ong, Medical Writer, Office of Clinical Sciences, Duke–NUS Medical School, for help in editing our article.

BIBLIOGRAPHY


Laryngoscope 128: July 2018 Kiong et al.: Meta-Analysis: IC as Selection Marker for CRT