Commentary: Statistical adjustment disorder: The limits of propensity scores

Paul J. Speicher, MD, MHS

For the past quarter century, anatomic lobectomy with nodal dissection has been accepted as the undisputed approach to early-stage lung cancer, based on the single prospective Lung Cancer Study Group trial. Over this time, however, much has changed regarding both the invasiveness—or rather, lack thereof—of our surgical approaches and the statistical techniques deployed to answer important clinical questions. What has not changed is the availability of prospective trials to better understand the topic at hand. Although many recent studies have appropriately called into question this long-held lobectomy dogma, nearly all are retrospective with varying levels of statistical sophistication.

The report by Tsutani and colleagues in this issue of the Journal is a valuable, albeit somewhat iterative, contribution that provides important insight into recurrence patterns following sublobar resection. Tsutani and colleagues examined 457 patients with clinical stage IA non–small cell lung cancer undergoing either wedge resection or segmentectomy, with the primary end point being recurrence. The authors rely on advanced propensity score techniques, rather than more traditional multivariable regression, as means of bolstering the impact of their study. This aligns with recent trends seen in the surgical and cancer literature globally, wherein propensity score–based methods are currently en vogue and have dramatically increased in recent years.

Unfortunately, although some guidelines do exist on appropriate use, including those we published a few years ago, propensity scores are often harnessed as a means of hand-waving away clinical factors and decisions that cannot be adjusted for. Tsutani and colleagues fall prey to this in some ways as well, attempting to use propensity techniques as a surrogate for surgeon decision-making and clinical expertise.

To that end, lymph node dissection or sampling is a fundamental tenant of lung resection for oncologic intent. The authors demonstrate a nodal recurrence rate among the wedge cohort, wherein nodes were not sampled at the time of surgery, of 7.7%. This is concordant with large population-based studies showing nodal upstaging rates of roughly 10% for clinical stage I disease following anatomic lobectomy. Interestingly, the opposing segmentectomy cohort experienced only a 4.2% nodal upstaging rate (2.7% at surgery and 1.5% regional recurrence in nodes). This suggests that segmentectomy was performed more frequently on those with perhaps less-aggressive tumor biology.

Similarly, there are significant differences in the use of a minimally invasive approach between the 2 groups, with 37.8% of segmentectomies and 27.2% of wedge resections performed open. This difference persists following propensity matching. Astute readers might question why so many wedge resections required thoracotomy within the last decade. Perhaps more importantly, with nearly 40% of segmentectomies performed via thoracotomy, no amount of statistical adjustment can address the proverbial elephant in the room: does open segmentectomy offer any meaningful benefit to the patient compared with video-assisted thoracoscopic surgery or robotic lobectomy, even in patients with borderline pulmonary function?

Ultimately, no amount of statistical adjustment will overcome intrinsic differences between groups, particularly attempting to use propensity techniques as a surrogate for surgeon decision-making and clinical expertise.
when the differences may have influenced the treatment choices themselves. Overall, this is a well-done study that certainly contributes to our understanding of sublobar resection for early-stage non–small cell lung cancer. However, it also serves as an important reminder of the intrinsic limitations of retrospective studies and advanced analytic techniques. Although statistical adjustment disorder may not appear in a future version of the Diagnostic and Statistical Manual of Mental Disorders, Tsutani and colleagues demonstrate that propensity score methods, as sophisticated as they can be, are neither a statistical panacea nor a replacement for prospective data.

References