Consensus-Derived Coronary Anastomotic Checklist Reveals Significant Variability Among Experts

Ara A. Vaporciyan, MD, MHPE, Vid Fikfak, MD, Matthew C. Lineberry, PhD, Yoon Soo Park, PhD, and Ara Tekian, PhD, MHPE

Department of Thoracic and Cardiovascular Surgery, Division of Surgery, MD Anderson Cancer Center, University of Texas, and Department of General Surgery, Houston Methodist Hospital, Houston, Texas; Department of Health Policy and Management, Zamierowksi Institute for Experiential Learning, University of Kansas Medical Center, Kansas City, Kansas; and Department of Medical Education, College of Medicine, University of Chicago, and Department of Medical Education, College of Medicine, University of Illinois at Chicago, Chicago, Illinois

Background. **Surgical skill assessment tools frequently reflect the opinions of small groups of surgeons. That raises concerns over their generalizability as well as their utilization when applied broadly. A Delphi approach could engage a broad group of experts to identify key elements for a checklist assessing coronary anastomotic skill, improving generalizability.**

Methods. **Expert surgeons in North America (10 or more years in practice, actively teaching coronary artery surgery) were contacted randomly to participate. Consenting surgeons first provided items they believed were mandatory when performing a coronary artery bypass. These were then entered into a three-round Delphi. Positive consensus was reached when 75% or more of participants ranked an item mandatory.**

Results. **Sixteen faculty consented to participate. Each participant provided 25 ± 10 items. The 407 items provided were condensed, resulting in 146 items in the final list, divided into six sections based on the conduct of the operation. Twenty-three items reached consensus in the first round, 14 in the second, and 3 in the third. These 40 items represented only 27% of the initial 146 items. Agreement within sections varied widely, from 0% for "management of assistants" to 47% for "testing and final steps."**

Conclusions. **A randomly selected group of experts using a Delphi approach can generate a checklist to assess construction of a coronary artery bypass. Considerable disagreement among experts regarding what steps are mandatory calls into question the generalizability of any locally developed checklist.**

(Ann Thorac Surg 2017;104:2087–92)

Education in cardiothoracic surgery, as with all surgical disciplines, requires the trainee to acquire both cognitive and motor skills. Whereas assessment of cognitive skills is addressed with a number of tools well suited to their evaluation (eg, multiple-choice question examinations), motor skills have very few assessment tools currently in use. To address this gap, new tools for many procedures have been developed, including checklists, objective structured assessments of technical skills, virtual reality simulators, and so forth [1, 2]. Unfortunately, important outcome measures are missed by many of these as the majority have been developed for basic surgical skills assessment, rather than advanced open surgical skills—the dominant skill in cardiothoracic surgery.

With roughly 400,000 CAB surgeries performed annually in the United States alone, coronary artery bypass (CAB) surgery is the most common procedure in cardiothoracic surgery [3]. A key component of that procedure is the construction of a CAB anastomosis, which is a multiple-step process. Considering that each CAB surgery requires an average of three anastomoses, roughly 1.2 million anastomoses are performed annually. In addition to its frequency, the procedure is also of short duration and can be easily simulated with both high and low fidelity simulators, making it an ideal procedure for assessment tool development.

Direct observation, which is the most commonly used method for assessment of surgical motor skills in CAB creation, suffers from poor reliability, poor compliance, and inaccuracy [4, 5]. Checklists and behaviorally anchored global assessments, originally created for

Accepted for publication July 17, 2017.

Presented at the Poster Session of the Fifty-third Annual Meeting of The Society of Thoracic Surgeons, Houston, TX, Jan 21–25, 2017. Winner of the Blue Ribbon as the top Cardiothoracic Education Poster.

Address correspondence to Dr Vaporciyan, Department of Thoracic and Cardiovascular Surgery, Division of Surgery, MD Anderson Cancer Center, University of Texas, 1515 Holcombe Blvd, Box 1489, Houston, TX 77030; email: avaporci@mdanderson.org.

The Supplemental Material can be viewed in the online version of this article [http://dx.doi.org/10.1016/j.athoracsur.2017.07.029] on http://www.annalsthoracicsurgery.org.

0003-4975/$36.00
http://dx.doi.org/10.1016/j.athoracsur.2017.07.029

objective structured assessments of technical skills [6], have been adapted for cardiothoracic surgery to improve the reliability and validity of direct observations of simulation exercises. The majority of these instruments, however, simply used the global rating component of the original objective structured assessments of technical skills tool and frequently dismissed the checklist component [7]. That was most likely the consequence of several studies showing that when compared with a checklist, the global rating scale provided better interstation reliability, better construct validity, and better concurrent validity [8, 9]. There was also no evidence that when added to the global rating component, checklists improved the reliability or validity of the global rating scale alone [10]. The limitation of all of these observational assessments tools, however, is that they lack reliability and validity at the specialist or higher trainee level [11] and exhibit a so-called "ceiling effect." That may be attributed to the design of the simulation models [12]; however, in assessment of live surgery, the lack of a model implies that any observed ceiling effect must be attributable to the assessment tool itself. Therefore, a more detailed checklist may increase its sensitivity and make it a more reliable and valid tool.

Another explanation for the limited sensitivity among existing checklists is the variability between institutions regarding the specific steps used. As highly complex procedures can be performed in a variety of ways with similar outcomes, single institution checklists may be populated with items that are institution-specific and detract from the overall sensitivity of the instrument. Unfortunately, the majority of checklists and GRS that have been developed lack broad participation [6, 13]. We hypothesized that a consensus building exercise that includes input from a randomly selected broad pool of clearly defined experts will produce a checklist that could address these needs. We intend to explore two specific elements of this hypothesis. First, is this approach feasible? And second, what is the degree of variability that exists between experts with regard to the items that eventually reach consensus?

Material and Methods

Selection of Consensus-Building Technique

After examining the existing consensus-building techniques, the Delphi method was found to be the best fit for our study owing to several unique characteristics. The entire Delphi can be done online, and therefore allows for global access to experts; the panel size requirements are modest, making the pool of experts queried manageable; and finally, the flexible design of the Delphi allows any number of follow-up interviews.

Selection of Experts

Experts were defined as North American cardiothoracic surgeons with at least 10 years' working experience after initial board certification, actively involved in performing coronary surgery, and teaching at an accredited cardiothoracic training program. A database of 314 North American experts who fit those criteria was created. Based on similar Delphi studies, 15 to 20 experts were sought for consensus building and were randomly invited to participate [14, 15]. When we had a minimum of 15 participants, the selection process was closed and no additional experts were invited to join. The final number of participants was 16 as more than one participant accepted the invitation simultaneously. Two participants were from the same institution, whereas the other 14 were from different institutions.

Delphi Method

Specific instructions were sent to selected participants asking them to provide items they believed were mandatory for competent performance of a CAB anastomosis. The instructions sent to each participant focused on four key elements. First, they were clearly instructed that any items they provided were "mandatory for the competent performance of a CAB anastomosis." These are the steps that "just have to be there otherwise it will not be a safe and well-constructed anastomosis." Second, the instructions defined what constituted a "CAB anastomosis," for example, when it began and when it was completed. Next, instructions on how to construct a checklist item were provided and included examples of well-constructed and poorly constructed items. Finally, to help organize the items and remind experts to address all aspects of a CAB anastomosis, the procedure was divided into six sections (dissection of the target vessel, creation of the arteriotomy, preparation of the graft, management of assistants, performance of the anastomosis, and testing or any additional manipulation of the graft). Experts were asked to provide as many items they believed were necessary for each section. As many as five reminders were sent to encourage submission of their initial self-created checklist.

When all the items were received, they were analyzed by the principal investigator and one uninvolved local expert, and similar items were grouped together. If any significant change in wording of the items was required, they would be sent back to the original participant to ensure that the true meaning of the item was preserved. Finally, after all revisions were approved, these condensed and edited items constituted a "master items list." This list was used for the Delphi consensus process.

In round one, all the items contained in the master items list were sent to the participants, who were asked to rank each item on a four-point scale (1 = not necessary, 2 = desirable, 3 = important, 4 = mandatory). After all the scores were obtained, a mean item score was calculated for each item. Consensus was defined separately for accepted and dropped items. We deemed that an item should be accepted (ie, "positive consensus") into the final checklist when 75% of experts ranked it as mandatory [3]. Conversely, an item was dropped (ie, "negative consensus") from the final checklist when its mean score was less than 2 ("desirable"). Items that fit neither of those criteria were advanced to round two.

In the second round, the participants were again asked to rank each item using the same four-point scale. This time, the items were accompanied with descriptive data derived from round one, including the minimum and

maximum, mean, variance, standard deviation, and score distribution (Fig 1). The participants were also now given an opportunity to comment on any item to either support it or prevent its inclusion in the final checklist.

When all the responses were obtained, the items that attained positive consensus were added to the final checklist, and items that attained negative consensus were dropped. The remaining items were sent into the third Delphi round, again accompanied with descriptive data derived from the second round along with anonymous participants' comments.

In the third round, the participants again voted on the remaining items, and when their votes were collected, items that achieved positive consensus were added to the accepted items from the first and second round and together comprised the final checklist. The rest of the items were dropped, and the study was terminated. Participants were advised that this was the final round and that items that did not reach a positive consensus would be dropped.

Experts' identities were blinded to all the participants throughout the entire Delphi trial and only visible to the principal investigator. Completion rate for all the surveys was 100%. All the data obtained in the study were stored in its original format for subsequent analysis.

Analysis

All responses from each round were analyzed using descriptive statistics.

Results

Expert Participants

A total of 100 faculty were contacted in just over 15 weeks. Four faculty revealed they were no longer active in coronary artery surgery. Of the remaining 96 experts, 25 (26%) agreed to participate and were provided with instructions. Nine participants subsequently either withdrew or did not submit a checklist despite five reminders. That left 16 expert participants who provided individual CAB anastomosis checklists and proceeded to the Delphi. The demographics of the expert participants and the entire pool of experts are shown in Supplement 1.

Initial Master Item List

Initial item generation by the participants required 1.6 ± 1.5 reminders. The time from the initial invitation to the provision of the item list was 31 ± 26 days (range, 0 to 89). The total time to collect all 16 participants' checklists was 22.6 weeks. The total number of items provided was 407 (Supplement 2). The average number ± SD and median number (range) of items submitted by each participant was 25 ± 10, 22 (11 to 48), respectively. During the generation of the expert-derived master list, it was noted that the "performance of the anastomosis" section of the CAB anastomosis contained items that were either very specific or more consistent with a general rule to be observed throughout the performance of the anastomosis. For example: "Toe sutures are placed as separate bites in the graft and the target" versus "Tissue handled gently to minimize tissue trauma." Therefore, a new section was created, titled "anastomotic general rules" to house the more general items pertaining to the performance of the anastomosis. The number of items suggested for each section and the final number after collapsing like items together into a master items list is shown in Table 1. The final number of unique items that constituted the master item list was 146. Of note was the rather consistent finding that two thirds of the items were duplicated by other participants in all the sections of the CAB anastomosis except for the "general rules" section, where only one third of items were duplicated by other participants.

Final Checklist Creation

The master item list was used for the consensus building exercise. The results of each round of the Delphi are shown in Table 2. Each round took roughly 10 weeks to collect all responses. After the first round, there were 23 items that reached consensus (15.8% of the master item list) and 52 items (35.6%) that were dropped. That left 71 items available for the second round. After the second round, an additional 14 items reached consensus and 5 were dropped, leaving 52 items for the third round. Only 3 items reached consensus in the third round.

Of the original 146 unique items included in the master items list, only 40 (27%) were identified as mandatory by consensus at the end of the three rounds of the Delphi exercise (Supplement 3). The majority of the items that reached consensus were identified in the first and second round (23 and 14 items, respectively, or 92.5%). Within each section of the CAB anastomosis, the percentage of items reaching consensus varied widely. The section with the greatest consensus was "testing and final steps," with 47% of the original 19 items reaching consensus. The section with the least consensus was "management of assistants," where none of the original 11 items reached consensus.

Provision of Comments

One hundred and nineteen comments were provided during the second round of the Delphi. Of the 71 items presented in the second round, 56 (78.9%) received a comment. When a comment was provided, the mean ± SD and median (range) were 2.1 ± 1.2 and 2 (1 to 5), respectively. Of the 52 items that were included in the third round, 40 (76.9%) items had at least one accompanying comment. For these, the mean ± SD and median (range) were 2.2 ± 1.2 and 2 (1 to 5), respectively.

Looking specifically at the participants, the number of items that received comments per participant ranged from none to 31. Eleven participants (69%) commented on at least one item. Overall, each participant commented on an average of 7.4 ± 9.3 items, or 10.4% of the items presented to them. Among the 11 who provided at least one comment, the average number of items commented on was 10.7 ± 9.5, or 15.1% of the items presented to them.

Comment

At the outset of this study, we hypothesized that a consensus building exercise could be applied to a random

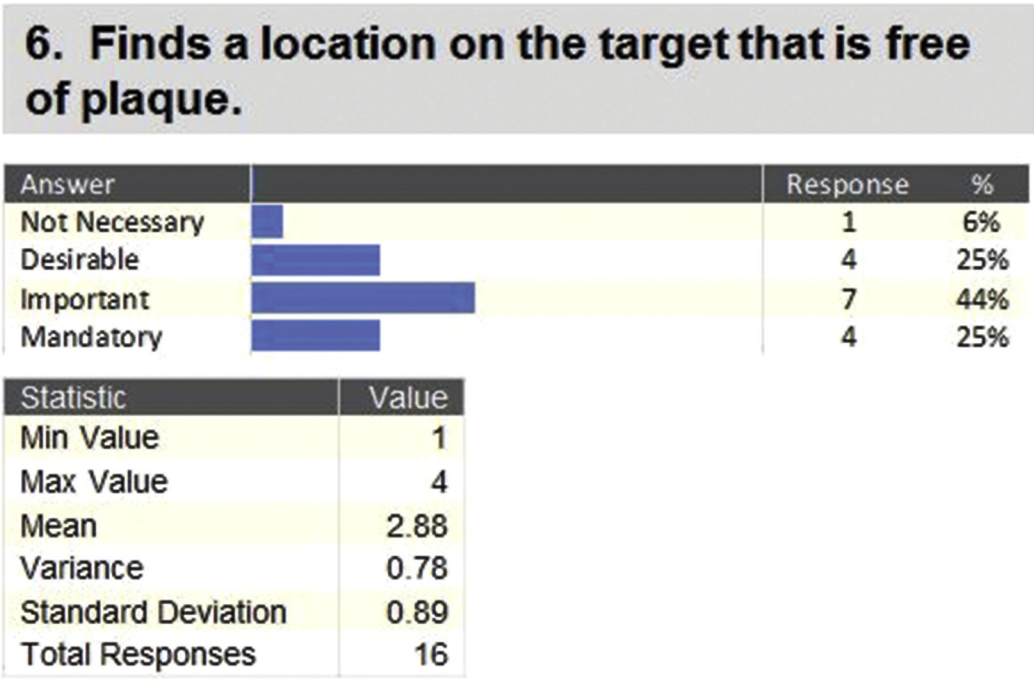

Fig 1. Example of an item presented to a Delphi participant, including score distribution. (Max = maximum; Min = minimum.)

EDUCATION

rational sample of experts to produce a checklist addressing the construction of a CAB anastomosis. The final product of this exercise was the production of a 40-item checklist that was created by 16 randomly selected experts across North America. Although the initial response was only 26% and the overall response rate 17%, considering the workload of an active cardiothoracic surgeon in a teaching institution, this number was higher than expected. Perhaps most representative of the feasibility of this approach was the robust participation of those who agreed to participate. The high mean and median numbers of items suggested by participants suggested that they took the task seriously. That there was significant overlap among the items suggested by participants (roughly 66% of items were repeatedly suggested by participants for all but one section) also reflects the high engagement of the participants. All the participants completed all three surveys with relatively limited prompting (only one or two reminders were ever needed). Finally, there was active provision of comments during the second round of the Delphi.

We were also able to address our research questions that focused on feasibility and variability among experts. The variability among participants was demonstrated by the high rate of item rejection during the Delphi. After the final round, only 27% of the original 146 items were considered, by consensus, to be mandatory despite clear instructions that the participants only provide items that they deem mandatory to the construction of a safe CAB anastomosis. The variability was greater in some sections of the procedure than others. For example, there was much greater consensus with items pertaining to "preparation of the vein graft," where half the items suggested reached consensus. Alternatively, none of the suggested

Table 1. Total Items Submitted by Participants and Final Numbers After Merger of Like Items

	No. Items Submitted by Participant			No. Items Before Merger	Items After Merger Into Master Item List	
Section	Mean	Median	Range		n	%
Dissection of targets	3.6	3	1–8	57	17	30
Creation of arteriotomy	3.7	3	1–8	59	20	34
Preparation of conduit	4.5	4.5	0–9	72	28	39
Management of assistants	1.9	1	0–7	31	11	36
Performance of anastomosis	6.8	7	1–18	108	37	37
General rules	1.1	1	0–4	18	11	61
Testing and final steps	3.9	4	0–8	62	19	31
Total	25.4	22	11–48	407	146	36

No. = number.

Table 2. Number of Items That Reached Consensus After Each Round

Section	Master List	Items That Reached Consensus			
		First Round	Second Round	Third Round	Final Consensus
Dissection of targets	17	4	2	1	7 (41%)
Creation of arteriotomy	20	2	0	0	2 (2%)
Preparation of conduit	28	7	3	1	11 (39%)
Management of assistants	11	0	0	0	0 (0%)
Performance of anastomosis	40	5	2	0	7 (18%)
General rules	11	1	2	1	4 (36%)
Testing and final steps	19	4	5	0	9 (47%)
Totals	146	23	14	3	40 (27%)

items in the section pertaining to "management of assistants" reached consensus. High degrees of variability are not specific to surgeons' technical approaches. For example, other investigators have found a similar high degree of variability in surgeons' preoperative decision making in distal pancreatectomy [16].

Unlike other reports utilizing a Delphi approach for checklist item development, our study differed in that all the initial items were created by the panel of experts, the expert panel was selected entirely at random (after applying clearly defined criteria for expertise), and consensus was clearly defined [17, 18]. These efforts were made to ensure that the final product would be generalizable across North America. Our use of experts from across North America revealed that the definition of what each surgeon considered a mandatory step varies widely. Nearly three of every four items thought to be mandatory by at least one expert participant could not reach consensus by the entire group. That calls into question the generalizability of any locally developed checklist assessing a procedure. Even more concerning, this variability was identified when analyzing one of the most common procedures performed by cardiothoracic surgeons. It would be safe to assume that more complex and technically demanding procedures (ie, coronary endarterectomy, mitral valve repair) would show even greater variability.

We can hypothesize two possible sources of this variability. The first may stem from some steps in a CAB anastomosis needing to be performed in a tight sequence. If one of these steps is rejected, then the other steps in the sequence must also be rejected. That may have contributed to some of the variability in this study, as evidenced by only 18% of the suggested items in "performance of the anastomosis" reaching consensus. This section is where the surgeons' actions would most conceivably be tightly sequenced. Conversely, items from sections that required less tightly sequenced steps, such as "anastomotic general rules" and "testing and manipulation of the graft," reached consensus at a higher rate. However, even if some sections of a CAB anastomosis require a sequence of steps, the experts still could not agree on even one sequence. The seven items that reached consensus in "performance of the anastomosis" are very general in nature and are not describing a tight sequence of steps.

The second source of variability could be that Delphi participants are unaware of what constitutes an essential step in a CAB anastomosis. That may be due to inexperience or lack of data. The lack of experience is unlikely as all the participants had at least 10 years of experience performing a CAB anastomosis. Lack of data is, therefore, a much more plausible explanation. Many procedural steps are based on a surgeon's prior training and the traditions and customs embedded within that training. Changing a technique that has been successful would carry some element of risk, and therefore, as long as the surgical outcomes remain excellent, these traditions are not discarded. However, that same blind preservation of each step makes it difficult for a surgeon to distinguish which steps are mandatory and which are based on tradition and could be adapted.

Limitations of this study should be noted. The response rate among the randomly selected participants was very low, as was expected owing to the significant effort required to provide an initial item list, then to complete three lengthy surveys. However, there will be a bias toward surgeons willing to engage in this activity and perhaps a bias against very clinically active surgeons who simply did not have the time available to participate. Another limitation may have been survey fatigue. The engagement of each participant over the many months required to complete all the parts of the study may have affected their commitment. Some participants may have simply passed judgment on items solely off the prior survey data that accompanied each item. Classically, this is the strength of the consensus-building activity. If a participant has any reservations about the value of an item, they should be swayed by the group's judgment and consensus will be reached. However, some participants may have been unwilling to dispute the group's judgment of an item not because of reservations but simply in an effort to complete the task or a perception that an additional survey would be needed if consensus was not reached. Certainly, we can hypothesize that this took place with some participants, but the use of a broad group of 16 participants should have minimized its impact. Additionally, the provision of comments in the second round by 69% of the participants suggests that these participants were still actively engaged in the process.

Our work has demonstrated that it is feasible to engage a broad coalition of randomly selected experts across a wide geographic area to produce a set of checklist items describing the mandatory steps in a complex procedure. The product created by this consensus-building exercise also revealed a wide degree of variability among participants in terms of what items are mandatory, which calls into question the generalizability of locally developed checklist that does not engage a broad group of experts. Future work will need to assess whether the checklist produced by this work can be used to assess trainees and faculty performing a CAB anastomosis and how Delphi-derived checklists compare with those created by other methods, such as hierarchical or cognitive task analysis, that use a limited number of experts.

References

1. Moorthy K, Munz Y, Sarker SK, et al. Objectives assessment of technical skills in surgery. BMJ 2003;327:1032–7.
2. Hamstra AJ, Dubrowski A. Effective training and assessment of surgical skills, and the correlates of performance. Surg Innov 2005;12:71–7.
3. Centers for Disease Control and Prevention (April 29, 2015). US Department of Health and Human Services. Available at http://www.cdc.gov/nchs/fastats/inpatient-surgery.htm. Accessed April 29, 2015.
4. Reznick RK. Teaching and testing technical skills. Am J Surg 1993;165:358–61.
5. Paisley AM, Baldwin PJ, Paterson-Brown S. Accuracy of medical staff assessment of trainees' operative performance. Med Teach 2005;27:634–8.
6. Reznick R, Regehr G, MacRae H, et al. Testing technical skill via an innovative "bench station" examination. Am J Surg 1996;172:226–30.
7. Lodge D, Grantcharov T. Training and assessment of technical skill and competency in cardiac surgery. Eur J Cardiothorac Surg 2011;39:287–94.
8. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med 1998;73:993–7.
9. Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ 2015;49:161–73.
10. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skills (OSATS) for surgical residents. Br J Surg 1997;84:273–8.
11. Ahmed K, Miskovic D, Darzi A. Observational tools for assessment of procedural skills: a systematic review. Am J Surg 2011;202:469–80.
12. Fann JI, Caffarelli AD, Georgette G, et al. Improvement in coronary anastomosis early in cardiothoracic surgical residency training: the boot camp experience. J Thorac Cardiovasc Surg 2008;136:1486–91.
13. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 2005;190:107–13.
14. Okoli C, Pawlowski SD. The Delphi method as a research tool: an example, design considerations and applications. Infor Manag 2004;42:15–29.
15. Nair R, Aggarwal R, Khanna D. Methods of formal consensus in classification/diagnostic criteria and guideline development. Semin Arthritis Rheum 2011;41:95–105.
16. Zilbert NR, St-Martin L, Regehr G, Gallinger S, Moulton CA. Planning to avoid trouble in the operating room: experts' formulation of the preoperative plan. J Surg Educ 2015;72:271–7.
17. Palter VN, MacRae HM, Grantcharov TP. Development of an objective evaluation tool to assess technical skill in laparoscopic colorectal surgery: a Delphi methodology. Am J Surg 2011;201:251–9.
18. Cheung JJH, Chen EW, Darani R, et al. The creation of an objective assessment tool for ultrasound-guided regional anesthesia using the Delphi method. Reg Anesth Pain Med 2012;37:329–33.