Evidence-based benchmarking in surgical performance: Leveraging the skill-outcome relationship in procedural assessment

Mitchell G. Goldenberg

Division of Urology, University of Toronto
International Centre for Surgical Safety (ICSS), St. Michael’s Hospital, Toronto

ICRE 2019
Ottawa, Ontario
Saturday September 28th, 2019
I have/had an affiliation (financial or otherwise) with a pharmaceutical, medical device or communications organization.

I am a paid scientific advisor for Surgical Safety Technologies Inc.
Assessing Performance in the Clinical Environment

1. Learning in the workplace is a linear, discrete process;
2. Competence is a fixed construct when abstracted from observed performance; and
3. Performance can be objectified based on observations made by assessors.

Challenged by Emerging Literature

1. Learning likely non-linear, especially in early learning curve, i.e. residency
2. Competence is contextual and task-specific; assessment validity is dynamic
3. Performance is hard to objectify, judgments are variable in clinical setting

---

Figure 1. Framework for clinical assessment.

---

Volume 65 • Number 9 • SEPTEMBER SUPPLEMENT 1990

2Louridas et al 2017, 3Varban et al 2016, 4Harris et al, 2017
Barriers & Challenges\textsuperscript{5}

Faculty *buy-in*

Variability in *standards*

Lack of faculty *training/orientation*

**Inadequate adjustment for clinical and patient parameters**

Case complexity may be a greater predictor of WBA scores than actual participant skill\textsuperscript{6}

Sampling Bias\textsuperscript{7}

\textsuperscript{5}Govaerts & Van Der Vleuten 2013, \textsuperscript{6}Wilkinson et al. 2008, \textsuperscript{7}Norcini et al. 2003
Assessment Validation in CBME

Kane’s Framework – *The Validity Argument*\(^9\)

- Scoring
  - Appropriate
- Generalization
  - Reproducibility
- Extrapolation
  - Alignment
- Implications/Decisions
  - Application

How do we *objectively determine* competency in technical performance?

What is the *benchmark*?

\(^9\) Kane et al. 1999
Setting Performance Standards in Technical Skills

Currently Used Methods of Standard Setting = Based-on Expert Judge’s **Conceptualization** of “Entrustability”, “Competency” or “Proficiency”

Predominantly *Simulation*-Based & Lack Exploration of *Consequences* of Standards

---

<table>
<thead>
<tr>
<th>Standard-setting method</th>
<th>Assessment setting</th>
<th>Type of assessment used</th>
<th>Total no. of studies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simulation</td>
<td>Clinical</td>
<td>Global rating</td>
</tr>
<tr>
<td>Participant-centred</td>
<td>18</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Contrasting groups</td>
<td>12</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Borderline group</td>
<td>1 (1)*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Generalized examinee-centred</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ROC curve</td>
<td>4 (1)*</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Item-centred</td>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Angoff</td>
<td>12 (1)*</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>+ Hofstee†</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>+ Ebel†</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>
Setting Performance Standards in Technical Skills

ONLY 16/37 included studies used judges who were:

1. Orientated
   Understand the purpose of the assessment
   Knowledge of the assessment tool used, including anchor points

2. Calibrated
   Demonstrate accuracy
   Inter-rater and intra-rater reliability

3. Content Expert
   Have knowledge of the task being assessed
   Familiarity with acceptable variations in performance

<table>
<thead>
<tr>
<th>Absolute method</th>
<th>No. of studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant-centred</td>
<td>n = 24</td>
</tr>
<tr>
<td>Content experts, trained in standard setting</td>
<td>8</td>
</tr>
<tr>
<td>Content experts, untrained in standard setting</td>
<td>4</td>
</tr>
<tr>
<td>Non-experts, trained in standard setting</td>
<td>1</td>
</tr>
<tr>
<td>VR simulator-generated metrics</td>
<td>5</td>
</tr>
<tr>
<td>Not specified in article</td>
<td>5</td>
</tr>
<tr>
<td>Item-centred</td>
<td>n = 13</td>
</tr>
<tr>
<td>Content experts, trained in standard setting</td>
<td>8</td>
</tr>
<tr>
<td>Content experts, untrained in standard setting</td>
<td>4</td>
</tr>
<tr>
<td>Non-experts, trained in standard setting</td>
<td>0</td>
</tr>
<tr>
<td>VR simulator-generated metrics</td>
<td>0</td>
</tr>
<tr>
<td>Not specified in article</td>
<td>1</td>
</tr>
</tbody>
</table>

10 Goldenberg et al 2016
Technical Performance in Robotic Surgery

**Global Rating Scales**

*Generic = GEARS*  
*Procedure-Specific = PACE*

**GEARS** = Strong validity evidence supporting its use across multiple assessment environments - *most ‘Ready for Prime Time’*

---

**Assessment method** | **Instrument description** | **Domains assessed** | **Number of studies, primary assessment method** | **Number of participants, primary assessment method** | **Number of studies, secondary assessment method** | **Number of studies, secondary assessment method** | **Content** | **Response process** | **Internal structure** | **Relationship to other variables** | **Consequences of testing**
--- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | ---
GEARS | Robotic-specific GRS; expansion of GOALS with expert consensus | Depth perception  
Bimanual dexterity  
Efficiency  
Force  
Sensitivity  
Autonomy  
Robotic control | 18 | 569 | 2 | 17 | 11 | 11 | IRR 0.38–0.92 (mean 0.80) | 1. Scores used to determine ranking  
2. GEARS score predicts surgical outcome | 12
PACE | Procedure-Specific GRS for RARP | Anchored Likert scale across seven operative steps | 1 | 56 | 0 | 1 | 1 | 1 | IRR 0.4–0.8 | 0 | 0

---

8. Goldenberg et al 2018
Hypotheses

1. Current methods of assessing technical skill in robotic surgery lack key sources of validity evidence that limit their use in high stakes assessments of technical skill.

2. Surgeon technical performance is a significant predictor of postoperative outcomes in robotic-assisted radical prostatectomy.

3. Standards can be set that use surgical outcomes to determine the benchmark, providing consequences validity evidence to assessments of technical performance in robotic surgery.
Objectives

I. Analyze intraoperative footage of RARP using **structured assessments** of robotic technical skill

II. Understand the correlative or predictive relationship between surgeon **technical performance** and clinically **significant outcomes** following RARP
Objectives

III. Use the predictive model regression equation to determine the performance score needed to predict a clinically significant outcome

IV. Use patient characteristics to adjust the assessment standard that reflects the heterogeneity seen in the clinical environment
Methodology

Multicenter Prospective Cohort Study

Intracorporeal Video Collected from consecutive RARP Cases

3 Content Expert, Trained Analysts

Surgical Steps Scored using GEARS and PACE

Three Primary Outcomes Selected

Erections @ 12m

Continence @ 3m

PSM
Results

Bivariate Analysis

Overall PACE Scores Significantly Higher in Patients with Erectile Function at 1 Year \( (p = 0.03) \)

Overall GEARS and PACE Scores Significantly Higher in Continent Patients \( (p < 0.01) \)

Overall PACE Scores Significantly Higher in Patients with Negative Margins \( (p = 0.02) \)

Steps Associated

Overall Only

Bladder Neck
NVB
Apical Dissection
UVA

Bladder Drop
Seminal Vesicles
Posterior Dissection
Apical Dissection
## Results

### Binary Regression Model

<table>
<thead>
<tr>
<th></th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall GEARS</td>
<td>3.53</td>
<td>1.01-12.43</td>
<td>&lt; .05</td>
</tr>
<tr>
<td>Age</td>
<td>0.98</td>
<td>0.92-1.05</td>
<td>.55</td>
</tr>
<tr>
<td>Nerve-Spare</td>
<td>0.89</td>
<td>0.21-3.75</td>
<td>.87</td>
</tr>
<tr>
<td>Volume</td>
<td>0.44</td>
<td>0.17-1.15</td>
<td>.09</td>
</tr>
<tr>
<td>BMI</td>
<td>0.83</td>
<td>0.31-2.24</td>
<td>.71</td>
</tr>
<tr>
<td>Posterior</td>
<td>1.82</td>
<td>0.66-5.03</td>
<td>.25</td>
</tr>
<tr>
<td>Reconstruction</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

Creation of *Composite Scores* Using β-Coefficients of Individual Predictive Steps

- **Bladder Neck GEARS Score**: 23%
- **Neurovascular Bundle and Prostatic Pedicle GEARS**: 17%
- **Apex GEARS**: 22%
- **Urethrovessical Anastomosis GEARS**: 38%

*Continence Composite GEARS Score*
Results

Creation of *Composite Scores* Using $\beta$-Coefficients of Individual Predictive Steps

**Continence Composite GEARS Score**

- Composite GEARS Score
- Overall GEARS
- Patient Age
- Posterior Reconstruction
- Prostate Volume $> 46.5$
- BMI $> 27.5$
- Nerve Sparing

**Odds Ratios and 95% Confidence Intervals**
<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>95% C.I.for EXP(B)</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total GEARS</td>
<td>-.606</td>
<td>.261</td>
<td>5.396</td>
<td>1</td>
<td>.020</td>
<td>.546</td>
<td>.327</td>
<td>.910</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.094</td>
<td>.057</td>
<td>2.773</td>
<td>1</td>
<td>.096</td>
<td>1.099</td>
<td>.983</td>
<td>1.227</td>
<td></td>
</tr>
<tr>
<td>BMI</td>
<td>-.019</td>
<td>.126</td>
<td>.023</td>
<td>1</td>
<td>.879</td>
<td>.981</td>
<td>.766</td>
<td>1.256</td>
<td></td>
</tr>
<tr>
<td>Prostate Weight</td>
<td>.006</td>
<td>.018</td>
<td>.106</td>
<td>1</td>
<td>.745</td>
<td>1.006</td>
<td>.971</td>
<td>1.042</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>6.257</td>
<td>6.641</td>
<td>.888</td>
<td>1</td>
<td>.346</td>
<td>521.741</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Positive if Greater Than or Equal To | Sensitivity | Specificity |
---|---|---|
0.0000000 | 1.000 | 0.000  
.0473255 | 1.000 | .043  
.2771744 | .958 | .391  
.2816372 | .958 | .435  
.3104328 | .958 | .478  
.3504087 | .958 | .522  
.3847808 | .917 | .522  
.4298695 | .875 | .522  
.4574342 | .875 | .565  
.8351015 | .042 | 1.000 
1.0000000 | 0.000 | 1.000  

ROC Curve

Goldenber et al 2018
Rearranged Regression Equation

$$GEARS = \frac{6.257 + 0.094\text{(Age)} - 0.019\text{(BMI)} + 0.006\text{(Prostate Weight)} - \ln[0.35-(1-0.35)]}{0.606}$$

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient age</td>
<td>62</td>
</tr>
<tr>
<td>BMI</td>
<td>28</td>
</tr>
<tr>
<td>Prostate weight</td>
<td>150</td>
</tr>
<tr>
<td>Constant</td>
<td>1</td>
</tr>
<tr>
<td>Probability of incontinence at 3 mo</td>
<td>35.0%</td>
</tr>
<tr>
<td>Total GEARS score needed</td>
<td>21.53</td>
</tr>
</tbody>
</table>

In order to provide a 62-year-old patient with a BMI of 28, a prostate size of 150 g with a 65% chance of being continent at 3 months postoperatively, a total GEARS score of 21.53 is required.
Results

Use of Weighted Scores to Set *Adjusted Standards* For Individual Procedure Steps and Overall

<table>
<thead>
<tr>
<th>Patient characteristics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>62</td>
</tr>
<tr>
<td>BMI</td>
<td>27</td>
</tr>
<tr>
<td>Nerve Spare</td>
<td>BILATERAL</td>
</tr>
<tr>
<td>VOLUME</td>
<td>46.5</td>
</tr>
<tr>
<td>STAGE</td>
<td>T2c</td>
</tr>
<tr>
<td>GLEASON</td>
<td>7</td>
</tr>
<tr>
<td>PSA</td>
<td>10</td>
</tr>
</tbody>
</table>

- **Probability of Continence at 3 Months**: 70.0%
  - Highest probability possible is 95.7%
- **Probability of Positive Surgical Margin**: 20.0%
  - Lowest probability possible is 5.5%

**Minimum composite score needed**

<table>
<thead>
<tr>
<th></th>
<th>for continence outcome</th>
<th>for PSM outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GEARS score</strong></td>
<td>4.03</td>
<td>4.05</td>
</tr>
<tr>
<td><strong>PACE score</strong></td>
<td>3.71</td>
<td></td>
</tr>
<tr>
<td><strong>PACE score</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

Use of Weighted Scores to Set *Adjusted Standards* For Individual Procedure Steps

<table>
<thead>
<tr>
<th>Minimum score by individual step</th>
<th>GEARS</th>
<th>PACE</th>
<th>PACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bladder Neck</td>
<td>3.69</td>
<td>2.07</td>
<td></td>
</tr>
<tr>
<td>Neurovascular Bundle &amp; Prostatic Pedicle</td>
<td>3.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apical Dissection</td>
<td>3.67</td>
<td></td>
<td>4.14</td>
</tr>
<tr>
<td>Urethrovescical Anastomosis</td>
<td>3.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Needle Entry (UVA)</td>
<td></td>
<td>3.39</td>
<td></td>
</tr>
<tr>
<td>Needle Driving (UVA)</td>
<td></td>
<td>3.28</td>
<td></td>
</tr>
<tr>
<td>Seminal Vesicle Dissection</td>
<td></td>
<td></td>
<td>4.19</td>
</tr>
<tr>
<td>Posterior Dissection</td>
<td></td>
<td></td>
<td>4.11</td>
</tr>
</tbody>
</table>
Key Findings

- Increase the acceptability of standards, based on clinical outcomes
- Adjustment for patient factors allows for comparative assessment between trainees/surgeons
- Novel Method of Weighting GEARs and PACE Scores Based on Clinical Significance
- Use as Tool to Predict Clinical Outcomes as well as Benchmarking Technique
- Potential use in credentialing/CPD of surgeons in practice
- Limit the use of surrogates for skill/performance
Thank You

PhD PAC Committee
  Dr. Teodor Grantcharov
  Dr. Antonio Finelli
  Dr. Jason Lee

Examiners
  Dr. Edward Matsumoto
  Dr. Darius Bägli

Collaborators
  Dr. Rajiv Singal
  Dr. Hossein Saadat
  Dr. Peter Szasz
  Dr. Tyler Hauer
  Dr. Alaina Garbens
Extra Slides
Assessments of Robotic Technical Skill

Task-Specific Checklists
Good internal structure
Easy to train raters, administer
Limited evidence for consequences validity evidence

Global Rating Scale
Excellent Internal Structure
More rater training required
Generic or procedure-specific

Safety Metrics
Commonly used in simulation-based assessment
FLS and FRS/FSRS

---

<table>
<thead>
<tr>
<th>Respect for Tissue</th>
<th>1</th>
<th>Frequently used unnecessary force on tissue or caused damage</th>
<th>4</th>
<th>Consistently handled tissues appropriately with minimal damage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time and Motion</td>
<td>1</td>
<td>Many unnecessary moves</td>
<td>4</td>
<td>Clear economy of movement and maximum efficiency</td>
</tr>
<tr>
<td>Instrument Handling</td>
<td>1</td>
<td>Repeatedly makes tentative or awkward moves with instruments by inappropriate use of instruments</td>
<td>4</td>
<td>Fluid moves with instruments and no awkwardness</td>
</tr>
<tr>
<td>Knowledge of Instruments</td>
<td>1</td>
<td>Frequently asked for wrong instruments or used inappropriate instrument</td>
<td>4</td>
<td>Obviously familiar with the instruments and their names</td>
</tr>
<tr>
<td>Flow of Procedure</td>
<td>1</td>
<td>Frequently stopped procedure and seemed unsure of next move</td>
<td>4</td>
<td>Obviously planned course of procedure with effortless flow from one move to the next</td>
</tr>
<tr>
<td>Use of Assistants</td>
<td>1</td>
<td>Consistently placed assistants poorly or failed to use assistants</td>
<td>4</td>
<td>Strategically used assistants to the best advantage at all times</td>
</tr>
<tr>
<td>Knowledge of Procedure</td>
<td>1</td>
<td>Deficient knowledge</td>
<td>4</td>
<td>Demonstrated familiarity with all aspects of operation</td>
</tr>
<tr>
<td>OVERALL PERFORMANCE</td>
<td>1</td>
<td>Very poor</td>
<td>4</td>
<td>Clearly superior</td>
</tr>
</tbody>
</table>

Goldenberg et al 2018
Results

Creation of Composite Scores Using β-Coefficients of Individual Predictive Steps

Continence Composite GEARS Score
- Urethrovessical Anastomosis GEARS 38%
- Bladder Neck GEARS 23%
- Neurovascular Bundle and Prostatic Pedicle GEARS 17%
- Apex GEARS 22%

Continence Composite PACE Score
- Needle Entry PACE 46%
- Needle Driving PACE 39%
- Bladder Neck PACE 15%
Results

Creation of *Composite Scores* Using $\beta$-Coefficients of Individual Predictive Steps

Positive Surgical Margins Composite PACE Score

- Apex PACE, 43%
- Seminal Vesicles PACE, 33%
- Posterior Dissection PACE, 24%
Results

Creation of Composite Scores Using $\beta$-Coefficients of Individual Predictive Steps

Continence Composite PACE Score

Odds Ratios and 95% Confidence Intervals

- Composite PACE Score
- Overall PACE
- Patient Age
- Posterior Reconstruction
- Prostate Volume > 46.5
- BMI > 27.5
- Nerve Sparing
Results

Creation of *Composite Scores* Using β-Coefficients of Individual Predictive Steps

Positive Surgical Margin Composite PACE Score

- Composite
- Overall PACE
- Stage ≤ T2b
- Stage T2c
- Stage ≥ T3a
- Gleason 6
- Gleason 7
- Gleason ≥ 8
- PSA < 4.5
- PSA 4.5-9
- PSA > 9

Odds Ratios and 95% Confidence Intervals
Key Limitations

Challenges around WBA’s in surgery
- Time and resource constraints
- Generalizability of assessment scores
- Attribution bias in skill-outcome relationship

Robotic Surgery in Canada
- Funding issues
- Limited evidence of benefit over open at population level
- Access to care and centralization

Non-Technical Skill Assessments
- Impact on technical skill
- Impact on patient outcomes
Ethical Considerations

Video Recording in the OR\textsuperscript{13}
Hawthorne Effect
Reporting and documentation of critical events

Data Security\textsuperscript{14}
Storage of Data
Classification of Data

Responsibility to the Patient
Impact of trainee participation in these studies
Reporting of ‘performance’ metrics

\textsuperscript{13}Langerman & Grantcharov 2017, \textsuperscript{14}Henken et al. 2012
Future Directions

Expanding the methodology

Inclusion of non-technical skills and human factors into model

Automation of skills assessment for real-time standards

Better understanding of surgical ‘complexity’ and how it impacts technical performance and outcomes

Application of this methodology across other procedural skills

Exploration of methodology in programmatic assessment

Prospective study of standard setting method *in training*

Prospective study of standard setting method *in credentialing*

Acceptability, feasibility and predictive ability of method using prospective data