

Apgar Scoring

From the recently-born to the recently-operated; simple scoring systems can go a long way for better decision-making.

**Rachelle Emard
Final Term Paper
Behavioral Decision Making:
Baruch Fischhoff
December 6, 2007**

Introduction

Doctors have to make important decisions every day. From diagnosis to treatment, they have the pressure to have the right answers, and they have the important task of making people's health problems better. There could be a thousand different variables that a doctor could take into account when approaching a problem, the uncertainties that exist when.

However, doctors are human beings, and have to work within certain constraints for decision-making. Through they may want to consider every variable, and probability, they have time constraints and are restricted to the variables they are able to reasonably measure. Doctors take vital statistics, have information about past medical history, and learn about different treatments and procedures; all in an effort to make sure they are able to make the best decision given their constraints.

Different decision-making heuristics, simplified rules-of-thumb, develop overtime to make decision making easier. Sometimes this can be detrimental, because human bias, such as overconfidence, availability and representativeness can skew our rational thought process. (Slovic) Still, people find their bias helpful because it makes their assessment easier. So ideally, for as far as medical evaluation is concerned, the best assessment system would need to be quickly measured, and have a worth-while probability of helping to assess risk.

This paper looks at the Apgar test and the development of an Apgar like test for surgery. I find the Apgar interesting because I see it as a case study for a simple measure can have a high degree of predictive power. I will also look at the benefits and shortcomings of doctors using the test to help direct diagnosis and treatment, from (a behavioral decision-making standpoint.

The History of the Apgar test

Virginia Apgar developed the Apgar test in the 1950's. Dr. Apgar developed the test in order to better assess the immediate health state of newborns in the perinatal period (right after). The test has 5 criteria, and each criterion has a ranking on a scale of 0 to 2. With all the criteria added together, the highest possible score would be a 10, the lowest being a zero. An Apgar score for a baby is taken 1 minute after birth, and then 5 minutes after birth. Depending on how low the Apgar score is, a doctor may want to take another Apgar score at 10 or 15 minutes. Though Dr. Apgar's measures involve some degree of subjectivity, it encouraged standardization of assessment, and was a big push in bringing evidence based medicine practices to the forefront.

Many doctors feel that the Apgar scoring systems revolutionized obstetrics. Saving many babies lives. The scoring scale is simple, and it has a component of qualitative assessment. 50 years later, the scoring system is the universal standard in almost all hospitals. Having this standard developed allows for doctors to better frame in their head what characteristics they are supposed to be looking for in an infant when they are born. A wave of critique came upon the scoring system in the 1980's doctors felt that there needed to be a more scientific way of assessing a baby's health, and they suggested testing the acidity of blood in the umbilical cord. In comparing the simple Apgar with the blood acid test, they found the Apgar to be eight times more accurate in predicting neonatal deaths. (Jacobson)

The test does have a degree of subjectivity, and I think that this portion is where the expertise of a doctor comes into play. A doctor knows what an “active” or “limp” baby looks like, and they know what that baby looks like in comparison to the average baby that they have seen in their experience. This subjective kind of expertise still has a role in medicine, the “art” part that Patterson talks about. There are things that linear models are not able to measure the same. Unlike a machine, a doctor can look at someone and realize that they are not well. Without the numbers, a machine just can’t compute that. I think that the Apgar for newborns works because it frames doctors to take some of their bias out of their subjectivity and really look at the specified criteria in order to get a sum assessment of a baby’s health at that point. It provides that they are looking at the same criteria after every delivery, and determining their next step from there. The value of the test comes in its ability to give immediate feedback about the condition of the baby, to the doctor.

Post Surgery Evaluation

Surgery carries a large amount of risk for the patient and the doctor operating. Every time you go under the knife, the patient has to realize that something could go wrong. The condition of the patient, and the outcome of the surgery is a crucial part of what a doctor would consider a successful surgery. Being able to measure the condition of a patient, and predict their risk of negative outcomes or complications is necessary in order to have an informed decision. The topic of this paper was sparked by a research paper I fell upon that was written by Atul Gawande, dealing with the development of an Apgar like score for assessing a patient after surgery. It hasn’t been put in prospective practice, but the retrospective results could show a lot of promise for the field. As of yet there isn’t a score that has brought to surgery what the Apgar has done for the treatment of newborns. Gawande looks looking to develop and use this Apgar for surgery assessment as a better predictor of future outcomes for the patient, instead of the surgeon going on their subjective expectations of how the operation will manifest itself.

Several surgical assessment scores/grades do already exist, though none of them are particularly effective or widespread. The simplest surgical risk assessment is the ASA grade. This grading system doesn't take into account any sort of actual measurement, and is a very subjective assessment of the patient, put into one of 5 different scales. This grade system is helpful for doctors, but it is too subjective to measure across doctors, hospitals, etc.

Another type of assessment is the APACHE surgical score which is a complex system of 14 categories used to predict outcomes for a patient. Where ASA is too subjective, the APACHE scoring system is too complex and intensive for a doctor to compute and quickly get the feedback for the surgery

Copeland developed what is thought of as being a compromise between the APACHE and ASA scoring system. His P-POSSUM score (Physiologic and Operative Severity Score for the enUmeration of Mortality and Morbidity), which was developed in the United Kingdom as a surgical audit procedure, has started to develop a following. The P-POSSUM score however, has a large amount of variables that are taken into account to figure it out. The scoring procedure involves a lot of common medical measurements such as age, the potassium levels, and pulse rate of the patients. However, you still have to take the number, multiply it by the weight of the variable in the POSSUM function, and then take the logarithmic function to determine the risk of mortality. It is generally found to be pretty accurate in predicting mortality in patients, but still the difficulty in calculation makes it impractical for immediate calculation in the operating room.

Developing an Apgar for Surgery

Though all of these measurements increase the predictability of outcomes better than a purely subjective statement of how well a surgery went, they still all lack the immediacy and impact that has made the Apgar so pervasive in it's industry. Gawande sought to figure out what easily measured variables could be found that kept the model and categories as simple as possible while still having

predictive power for risk assessment. This is important not only because it would be easier for the doctor to calculate, but because it would make it easier for the non-expert to comprehend and understand (patients, family members)

How do you go about creating an Apgar like scoring system for surgery? Perhaps Virginia Apgar just got lucky with her choice of characteristics, and her scaling system. But that's unlikely. To develop the score, you would have to look at the metrics in a surgery that are available to you, and determine which of these criteria are most important, or influential to whether the patient has a positive outcome. Like in creating a linear model, it involves finding which characteristics are most important at that point in the patient's journey. Gawande looked at 17 different potential predictors, and then created and tested three different Apgar-like models to surgical data, looking for what was most predictive. He tested the first cohort of patients retrospectively, where they involved patients that had only gone through colonoscopy, and then created a model from that cohort, and applied it prospectively to a broader range of surgical operations.

The result was easily calculable formula based on three characteristics of the surgery patient. He tested three different models or three different cohorts. The resulting model is a 10 point score that measures 3 characteristics: Estimated blood loss, lowest heart rate, and lowest mean arterial pressure. (Gawande) The lowest heart rate is able to reach 4 points instead of 3, so it is weighted slightly higher, though it is still easy to calculate in your head. (Appendix 2)

The success of Apgar also comes from a refocus of attention. Before the Apgar for newborns, many doctors would focus on the immediate condition of the mother that had just given birth. With the Apgar they were forced to take notice of similar characteristics in all babies. Typically, doctors still have to crutch on their subjective opinion about the status of their post-operative patient, it could be right, but chances are they are affected by the human biases that we all experience in our lives.

The importance of an Apgar like measurement after surgery is that it can really help doctors assess how a surgery went for the patient, and how their state after surgery went as compared to what complications (if any) were expected to be like. Surgeons typically have follow-ups with the patients after the procedure to assess the success after the surgery. At this point they are prone to a confirmation bias about the procedure. As discussed in the Deyo article doctors and patients are prone to confirm more positive results, because they want to believe that their state is good, they of course want to hear that the surgery went well, that they are doing well based on what they thought would happen after the surgery. If the patient is alive and healthy it is because the surgery went well. If the follow-up does not go well, the surgeon can still establish whatever subjective reason for why that was so. With an accurate and easily calculated benchmark, post-surgery to help assess 30 day risk, they can go in with that in hand.

One issue with the process of development of the Apgar for surgery is that it was developed using retrospective analysis. It is the hope that people would continue to collect prospective data after surgery in the same way, but perhaps they would not get the same results, though Gawande believes that the scoring system is replicable. More importantly, if the surgery Apgar score is utilized so the doctor is able to give better treatment, then than the predicted positive outcomes over time should increase. Making overall surgery risk of complications decrease

Of course, any kind of surgery is a risky endeavor. One score should not be the be-all, and end-all of post-surgical decision making, because like most big decisions, there are many different variables involved. I think that the Apgar for surgery could be a great component and tool, when paired with a doctors subjective feelings about the patient's status. Interestingly, where the newborn Apgar has a lot to do physical, visual, and sensory components, the surgery Apgar is void of that. It gets its metrics from very basic operating room stats. Which means that the subjective component is still brought to the

table by the doctors. The Apgar can't "see" the patient, and these physical components I feel would also be important in getting the big picture of the situation and making a decision based on those results.

I have focused a lot on a doctor's decision making, but they are not the only decision makers in the picture. A patient's family members are typically in the picture as well, and they can be put in the position to make a big decision about the treatment of their loved ones. If you are able to give a non-expert, a score that they can understand, that can still have predictive power, it will help the family members better grasp the decision they are about to make and the risk associated with it. If using the Apgar can be very helpful in this capacity, when a patient goes into surgery and comes out they could have an understanding of their state. Concepts like blood loss, and heart rate, are easy to understand, whereas if someone told me my potassium is a concern, I wouldn't know exactly why potassium in particular would have a impact on my health, I wouldn't know what my potassium level has to do with my surgery. The doctor could break down my entire POSSUM score, but it would take a lot of understanding to really know what it would mean.

Where the Apgar tests (both for surgery and for newborns) is *not* a helpful measure is in the comparison or evaluation of a doctor, hospital, floor, etc. You can't compare performance really between different patients either. The score is almost exclusively a measure of the physical health of an individual patient as it relates to the anticipated level of risk that they have in having complications after surgery. The doctor will not be able to assess "how good of a doctor he/she is" based on their track record of surgery Apgars. The only thing that it would tell them is about the average health assessment of the doctor's patients after they had surgery.

Behavioral Decision Making Connection

Babies, surgery, medicine, are all swarmed with important decisions, where we differ is in our processing and interpretations of these decisions. What can we do to make the decisions easier? Why aren't we the optimal decision-maker?

Dawes' discussion about improper linear models can certainly be applied here. The more complicated versions of the surgery evaluation scores, such as the POSSUM and APACHE, have very many variables, and have been weighted and calibrated in order to get the highest correlation with outcomes. It is important to have models in post-surgery evaluation. As good as a surgeon is, as many surgeries that he/she has completed, the truth remains that humans are generally crummy at being able to accurately assess between multiple variables in their head. The great thing about the Apgar scores that were developed is that it keeps it simple. The simpler, the better, as long as you are not seriously sacrificing accuracy. In doing research I stumbled upon "Apgar" scores that people created for other things. "Apgar" for your finances, "Apgar" for class involvement, an "Apgar" for software development. They all are essentially embracing and celebrating Virginia Apgar's use of a simple, and easy to calculate model, weighing them equally, to help you assess a situation, problem. The result, as Apgar has evidence, and as Dawes has shown is that you are able to make better-informed decisions, cutting back on problems of availability, and representativeness, and overconfidence.

Comparison is definitely involved with the Apgar, just as school grades induce people to compare themselves to others. Where comparison would not be accurate here is in the comparison of the doctor, and the comparison of hospitals, etc. Just as with the Apgar for newborns, you wouldn't be wise to pick a hospital that has a higher average Apgar score for their deliveries, because it has to do more with the nuances and condition of the individual baby, rather than the quality of care at the hospital. Where someone could compare is between their individual Apgar scores, realizing what their health was after an operation as compared to what their health was after a prior operation. Still, they couldn't do much

to change or improve their Apgar score, but they could realize what their state of health is and know what their risk looks like as compared to the last operation they had. Regardless of the outcome, the development of a realistic, and consistent comparison tool is more helpful than having a doctor give you input that was possibly tainted by his most recent run of surgeries. (Schwartz)

Establishing risk post operation establishes certain expectations for the patient and the doctor. If they out-perform their expectations, say they had a 60% risk of complications, and had a wonderful recovery, and then the doctor and the patient will be excited with their turnout, feeling that their outcome was better than their expectation. Say they “under-perform” based on expectations. Then the doctor and the patient would also be surprised, and less satisfied. It is important to realize that what the Apgar would be measuring is not the surgery experience, but they are measuring their state of health.

Implementing the scoring system also bolsters Patterson’s arguments for evidence medicine, and in the Newborn Apgar’s case, it is an example of when the nuances of subjectivity can have a place in medicine. On another end, I feel like the Apgar type of test, especially as it relates to newborns is a testament to the balance using a little bit of a doctor’s know-how and tacit knowledge instead of just looking exclusively. In Patterson’s article, “What doctors don’t know” (almost everything) he pushes for evidence-based medicine; he feels that with good intentions, it could be the hierarchy-busting revolution that patients and doctors have needed. He talks about the expert cardiologist and the computer’s superiority in catching the signals for diagnosis. The Apgar is successful because it does allow for the doctor or nurse to use a level of subjectivity. What is limp? What is almost no crying? What constitutes a sneeze, should below average sneezes count? Doctors and nurses have seen lots of babies, and are keen on each of these categories. They likely don’t need to think much about exactly what score to give for each category, because they recognize the norm.

In conclusion, models and scoring can be very helpful in combating our human errors. When developing a model the old KISS rule (Keep it Simple Stupid), when it can be applied, should be applied. This brings the more calculated strategy on par with the short-cut heuristics that we tend to use. No one wants to be complicated, but everyone wants the best answer. Also, the development and success of the Apgar shows how these systems can contribute to better-guided decision-making from the consistency that it offers doctors and patients, and in the case of the Apgar for newborns, still be effective, still making a difference fifty years down the line.

Appendix 1-Apgar for Newborns

Apgar Criteria	Score of 0	Score of 1	Score of 2	Acronym
Skin color	blue all over	blue at extremities body pink	no blue cyanosis body and extremities pink	Appearance
Heart rate	absent	<100	>100	Pulse
Reflex	no response to stimulation	grimace/feeble cry when stimulated	sneeze/cough/pulls away when stimulated	Grimace
Muscle tone	none	some flexion	active movement	Activity
Respiration	absent	weak or irregular	strong	Respiration

Appendix 2 Proposed model for Apgar for Surgery

Table 4. A 10-Point Surgical Outcomes Score*

	0 points	1 point	2 points	3 points	4 points
Estimated blood loss (mL)	> 1,000	601–1,000	101–600	≤ 100	—
Lowest mean arterial pressure (mmHg)	< 40	40–54	55–69	≥ 70	—
Lowest heart rate (beats/min)	> 85	76–85	66–75	56–65	≤ 55 [†]

Surgical score = sum of the points for each category in the course of a procedure.

*Based on model 1 from cohort 1.

[†]Occurrence of pathologic bradyarrhythmia, including sinus arrest, atrioventricular block or dissociation, junctional or ventricular escape rhythms, and asystole also receive 0 pts for lowest heart rate.

Bibliography

- Dawes, Robyn. "The Robust Beauty of Improper Linear Models in Decision Making." American Psychologist 34.7 (1979): 571-582.
- Deyo, Richard. "Practicing Variations, Treatment Fads, Rising Disability ." Spine 18.15 (1993): 2153-2161.
- Gawande, Atul , Mary Kwaan, and Scott Regenbogen. "An Apgar Score for Surgery." American College of Surgeons 204 (2007): 201-208.
- Gawande, Atul . " Medical Dispatch, "No Mistake," ." The New Yorker 30 Mar. 1998: 74-80.
- Gawande, Atul. "The Score." The New Yorker 9 Oct. 2006. 1 Dec. 2007
<http://www.newyorker.com/archive/2006/10/09/061009fa_fact>.
- Jacobson, Sherry. "Apgar Test Still Scores As Neonatal Predictor." The Washington Post 1 Mar. 2001. 1 Dec. 2008
<<http://pqasb.pqarchiver.com/washingtonpost/access/69210542.html?dids=69210542:69210542&FMT=ABS&FMTS=ABS:FT&date=MAR+04%2C+2001&author=Sherry+Jacobson&pub=The+Washington+Post&desc=Apgar+Test+Still+Scores+As+Neonatal+Predictor%3B+System+Tops+Newer+Methods+in>>>.
- Ludwig. "Virginia Apgar (1909-1974)." Der Gyn 40.3 (2007): 227-228.
- "POSSUM - Background Information - www.riskprediction.org.uk." Risk Prediction in Surgery - www.riskprediction.org.uk. 6 Dec. 2007
<<http://www.riskprediction.org.uk/background.php>>.
- Patterson, Kevin. "What Doctors Don't Know (Almost Everything)." New York Times Magazine 5 May 2002: 77-79.
- Schwartz, Barry. The Paradox of Choice: Why More Is Less. New York: Harper Perennial, 2005.
- Slovic, Paul, Baruch Fischhoff, and Sarah Lichtenstein. "Rating the Risks." Environment , 21: 14-39.
- "The Virginia Apgar Papers: Obstetric Anesthesia and a Scorecard for Newborns, 1949-1958." Profiles in Science. 6 Dec. 2007
<<http://profiles.nlm.nih.gov/CP/Views/Exhibit/narrative/obstetric.html>>.