Reviewer: 1

Comments for the Author
Bhangu and colleagues have submitted an extremely interesting paper, employing a creative technique of capturing very difficult and previously undercollected and unreported data on mortality following abdominal surgery. Using an online platform, they report on the outcomes of 10,745 patients undergoing emergency abdominal surgery from 357 facilities in 58 countries. Data were entered into an electronic database (REDCap) by collaborators at each facility for two consecutive weeks during the period from July-December 2014, essentially crowdsourcing the data collection effort. They stratified countries based on Human Development Index (HDI), and report on crude mortality and morbidity rates as well as rates after adjusting for patient characteristics and facility and service characteristics.

They report an overall 24 hour crude mortality rate of 1.8%, which increased to 5.3% after 30 days (or hospital discharge). Mortality rates increased across HDI country groups at both 24 hours and 30 days and increased after adjusting for patient factors, although this seemed to disappear (or at least approach baseline) after adjusting for facility level factors.

Regarding the statistical techniques, strategy, and methodology, I am not a statistician but do wonder what was done to control for clustering within facilities – the authors note an attempt to adjust for country-level variation using random effects, but I'm not sure this is adequate given that 22 of the 58 countries only supplied data from a single center and clustering by facility is clearly going to be an issue.

I have a number of serious concerns that may prove difficult to address based on the methods described in the paper and the protocol.

1) Was there a method for confirming, or independently auditing, any of the data that were being submitted through REDCap? I ask because I am concerned that without such audit, there will always in my mind be a question of data fidelity. While this effort was admirable and remarkable, it is still fairly new and there needs to be some way to validate reported data.

2) What was the process by which centers were recruited to the study? In the protocol, there is mention of a one-day data entry "trail" run using a prestudy survey (the WHO situational analysis tool). How was that used? Were the same individuals entering these data also entering data for the two weeks of the study and if not, what was the purpose of collecting such data? Were centers removed from the study based on this prestudy work? Did the prestudy work allow for training, and if so, how, and what form did this training take?

3) How are differences between HDI countries treated with respect to time from admission to surgery? Acute abdominal operations usually take place in the context of a time-dependent disease state. Delays from presentation to operation potentially increase morbidity and mortality. In table 1 there is a description of the differences in timing of operation - 17.7% of operations occurred 24 hours or more after admission in low HDI countries, compared to 37.4% in high HDI countries. To my mind this indicates a fundamental difference in presentation, diagnoses, and clinical decision making, as high HDI countries presumably have all appropriate human and infrastructural resources at their disposal. With such assumptions, then these collected cases and conditions may not be comparable across HDI strata, as the findings are, in fact, contrary to what I might have empirically expected – that high HDI settings would move quickly to surgical intervention for patients requiring emergency abdominal operations. Given that the findings demonstrate the opposite, the most obvious explanation is that these cases, despite their presumed "urgency", are indeed different.

4) Which begs the question, are these patients and diagnoses comparable? After reviewing the supplementary materials, the

summary diagnosis table would indicate there are profound differences in the case mix by diagnosis alone. How was this handled? The methods are not clear on this, and need to be greatly expanded, either in the supplementary materials or the text itself, to explain how differences in presentation, diagnosis, and operation performed are handled and adjusted to allow comparisons across HDI settings.

5) In addition, while the range of operations has been limited to emergency abdominal surgery, this is still a very heterogeneous set of procedures. There is a big difference between a laparoscopic cholecystectomy or appendectomy for uncomplicated disease, an exploratory laparotomy with bowel resection for dead gut, and repair/resection of bowel for perforation. The Pearse study quoted by the authors attempted to adjust for the complexity of the operations; in this analysis, however, there does not seem to be any attempts to adjust for the type, complexity, or inherent danger of the procedure itself.

6) I suspect that there are major differences between LMIC facilities that report data in this study and other facilities within the same countries that have not or could not. Do the authors have a sense of how representative such facilities are of the general infrastructure for surgical care in these countries? What the authors report is essentially a convenience sample, which is not in itself a problem, but this needs to be recognized more candidly in the discussion.

Other issues, points, concerns, and recommendations:

1) Page 6, second paragraph – this needs to be more fully described, in particular the "diagnostic categories" (which I assume refers to "diagnostic type" in table 1), and "service variables" (which I assume refers to the last five variables listed in table 1).

2) Page 6, line 45 – referring to the two sets of regression models, is the baseline patient characteristics the first adjustment (as labeled "baseline" in figure 2)? Does the second regression model just include adjustments for measures of hospital facilities and services, or does it also include the prior adjustment for patient

characteristics (presumably labeled “full” in figure 2)? If the latter, what does the model look like when adjusting for facility and service characteristics alone – how much does that affect the unadjusted results?

3) Page 6, line 47 – what are the hospital facility and service measures that are being used in the adjustment model? I assume they include surgeon/anesthetist experience, checklist use, type of anesthetic, oxygen, pulse oximetry, prophylactic antibiotics, whole blood, and thromboembolic prophylaxis, but I cannot tell from this description in the methods.

4) Page 9, line 35 – if no ICU facility exists, how does this change practice across HDI settings with respect to “requiring” ICU admission?

5) Page 12, line 14 – there is no reference 19, please update.

6) Page 12, line 40 – this is reference 9, I believe, not reference 1.

7) Why was the protocol not translated into Arabic (page 23, line 18)? Many of your participant countries would have found this useful and 56 of your 357 centers were based in countries where Arabic is primarily spoken.

8) Table 1 – regarding oxygen, pulse oximetry, prophylactic antibiotics, whole blood, and thromboembolic prophylaxis as listed at the bottom of this table, do these refer to their presence and availability, or actual use? I would expect some combination, so please clarify. In particular, the use of prophylactic antibiotics for emergency abdominal surgery may actually be a mute point, particularly for infections or bowel perforation, as contamination in these circumstances is already established and therefore no “prophylaxis” occurs. This would probably be best broken out into a separate table. In fact, separating table 1 into at least two tables – one for patient demographics and another for infrastructural/service/facility variables and characteristics might help make the data more understandable.

9) Table 2 – it is currently listed as table 4. This too might best be broken into two tables as outlined in #7 above. As it stands, I cannot determine a clear message or set of conclusions from these results.

With regard to the fundamental assumptions underlying this study, I would hypothesize (but have limited empiric evidence) that if excessive mortality in LMICs following fairly routine, uncomplicated surgery exists, it is driven primarily by 1) extremely delayed presentation, 2) high infectious complications, 3) different prevalent comorbidities such as malnutrition, parasitic coinfection, or HIV, and 4) high rates of failure-to-rescue (that is, complications otherwise manageable with critical care infrastructure and aggressive support/intervention result in death). On the other hand there is the potential for selection bias in any setting, but particularly in weak health systems where surgeons (and facilities more generally) may not operate on high risk patients; deaths following surgery reflect poorly on the facility in the eyes of the community regardless of cause or preventability and this skews postoperative mortality results since such patients never even get the intervention in the first place. Under these circumstances perioperative mortality may be lower in such settings. The results of this study do not clarify these issues.

There are lots of data in the extensive tables, but I still do not have a clear sense of what all this information means. In particular, the adjustment by both patient variables and then facility and service variables does not provide a clear sense of what drives the variability in mortality. I think the authors are noting that service and facility variation accounts for a high proportion of this variability (page 12, 3rd paragraph), but this conclusion is not entirely evident from the data.

Reviewer: 2

Comments for the Author
The authors present a clearly written study of international outcomes following emergency abdominal surgery. I have several questions and comments.

1. The grouping of hospitals into terciles of HDI ignores the many inherent differences in the individual hospitals included. Was there a subset analysis performed within each tercile to see if there are high or low performing hospitals within the terciles and what those hospitals looked like?

2. There is significant collinearity in the various data points and outcomes presented. For example, appendectomy was the most common "emergency" operation in the high HDI group and trauma was the highest in the low HDI group. It is therefore fairly obvious that the mortality rate in the high HDI group would be lower at baseline. The subset analyses are helpful (i.e., appendectomy results similar across HDI groups), but only further strengthen the finding that low HDI hospitals are doing fine with most emergency operations, like appendectomy. The mortality following laparotomy was higher in low/middle HDI groups, but not the 2-3 fold that the paper describes overall.

3. There is a good amount of literature supporting the finding that complications are a reasonable proxy of patient severity and mortality or failure to rescue are better proxy's for hospital level quality. This is an interesting finding in the authors' data that does not receive much attention in the manuscript. The complication rate in the high HDI hospitals was higher across the board, but they had allowed case-fatality or failure to rescue rate. This helps focus the attention of the institutions on the timely recognition and effective management of those complications. The authors should expound upon this further.

4. The finding that more experienced anesthetists have worse outcomes doesn't make sense and the explanation presented doesn't account for the differences. The authors should make a critical appraisal of that finding. Perhaps they could evaluate the severity of illness in the cases those anesthetists are involved with.

5. While I applaud the authors on the establishment of this international collaborative and data registry, I disagree with their conclusion that "The marked variations reported indicate the need for wider implementation of quality improvement initiatives to address the wide disparities in outcome." The authors fail to mention throughout the manuscript the obvious resource deficiencies that the low HDI hospitals may inherently have. Therefore, while a quality initiative might be to increase the utilization of checklists, that will likely have little to no benefit to a low HDI hospital that can't afford laparoscopic or advanced interventional radiology equipment. There are basic necessities that may be lacking and would require substantial investment before those hospitals try to improve adherence to first world quality improvement interventions. Finally, quality improvement is a local process. The solutions at each site are going to be very different given the vast heterogeneity in hospital types included in this study.

Thank you for the opportunity to review this manuscript.

Reviewer: 3

Comments for the Author
This manuscript documents a remarkable multinational effort to measure outcomes of a highly morbid set of procedures across a spectrum of Human Development Index countries. However, what is remarkable is how they pulled it off, not what they found. The differences in outcomes that the authors describe are fairly predictable in terms of direction and magnitude. If, however, the results were not in line with my a priori assumptions, then I would not be inclined to believe the results because of methodological weaknesses of the study: the authors fail to describe how they came up with the sample of hospitals. One must assume it was a convenience sample. A convenience sample might be acceptable if the intent of the study were to demonstrate feasibility of outcome measurement across the world, but it is not acceptable if the aim is

to measure the magnitude of the differences in outcomes across countries of different wealth. This manuscript would be more appropriately cast as a feasibility pilot study based on its methodology. The content of the manuscript would then more appropriately focus on the process of conducting the study, obstacles overcome, and lessons learned.