storing their handguns loaded and unlocked during the subsequent bedside interview (p=0.10). The 6 patients (10.2%) who reported storing long guns loaded and unlocked and patients who stored firearms in other ways were generally consistent in their answers between survey methods

Conclusion: In ED patients asked about firearms, some responses to a self-administered survey modified from a previously validated tool were inconsistent with verbal answers obtained from bedside interview. This inconsistency could be a challenge for HCPs and researchers who rely on the accuracy of patients' self-reported answers, and potentially raises questions about the results from previous research in the field . Stakeholders who utilize similar survey question-types should consider including a verbal confirmation to ensure the accuracy of responses.

412 Testing and Enhancing an Emergency Department Acute Heart Failure Risk Stratification Tool

Dana R. Sax[1], Dustin G. Mark[1], Jie Huang[2], Jamal S. Rana[1], Sean P. Collins[3], Alan B. Storrow[4], Dandan Liu[4], and Mary E. Reed[5]
[1]*The Permanente Medical Group,* [2]*Division of Research, Kaiser Permanente Northern California,* [3]*Vanderbilt University Medical Center,* [4]*Vanderbilt University,* [5]*Kaiser Permanente Division of Research*

Background and Objectives: There are over 1 million ED visits per year for acute heart failure (AHF), and over 80% of patients are admitted. Several recently validated ED-based AHF risk stratification tools have focused on identifying lower risk patients who may be eligible for discharge. We used one of these rules (the STRATIFY rule, which includes 13 variables) as a basis for further optimization of risk prediction using additional patient history and clinical data

Methods: We identified all adult health plan members of a large integrated delivery system who had an ED visit for AHF between 1/1/2017 and 12/31/2018. The primary outcome was any 30-day serious adverse event (SAE), including death, cardiopulmonary resuscitation, intra-aorta balloon pump, endotracheal intubation, new dialysis, myocardial infarction or coronary revascularization. We first applied the STRATIFY rule, and then tested whether predictive accuracy could be enhanced in two ways: 1) by testing use of additional variables (approximately 70, available within 3 hours of ED arrival) in logistic regression and 2) by using additional variables and machine learning models including decision tree, random forest and XGBoost. We divided the sample into training (80%) and testing (20%) cohorts and assessed performance among the testing cohort. We reported area under curve (AUC) of logistic regression and the best performing machine-learning model using all variables and assessed admission rates among several risk categories

Results: There were 26,189 ED encounters; mean age was 74, 51.7% were women and 60.7% were white. The overall 30-day SAE rate was 18.8%. The SAE rate among discharged, observed, and admitted patients was 5.7%, 11.0%, and 26.2% respectively. The AUC of the STRATIFY rule was 0.71. With additional variables, the AUC was 0.80 from logistic regression and 0.85 from XGBoost. We found that 11.1%, 25.7%, and 48.9% of our population had of ≤3%, ≤5%, and ≤ 10% SAE risk (predicted from XGBoost), respectively. Admission decisions correlated with risk, although over 27% of patients with a ≤3% risk were admitted, and 18% of patients with a ≥10% risk were discharged

Conclusion: Use of a machine learning model with additional variables significantly improved predictive accuracy and if easily implemented into ED workflow may help to safely decrease hospital admissions among lower-risk patients.

413 Diagnostic Tests in Emergency Medicine Journals Under-Report Confidence Intervals and Likelihood Ratios

Ari B. Friedman[1], Aric Berning[2], and Keith A. Marill[3]
[1]*Penn Medicine-Department of Emergency Medicine,* [2]*University of Pittsburgh,* [3]*Harvard Medical School/Massachusetts General Hospital*

Background and Objectives: 95% confidence intervals (CI) and likelihood ratios are crucial for the appropriate interpretation of diagnostic tests and the uncertainty surrounding an estimate. We hypothesized that confidence intervals around sensitivity and specificity would be reported nearly always, that likelihood ratios would be imperfectly reported, and that likelihood ratio confidence intervals would be rarely reported

Methods: We searched Pubmed or Scopus for all articles reporting diagnostic test results in 20 top EM journals from 2011-2016, and randomly sampled 124 articles. Sampled articles were abstracted by trained researchers with high agreement (intraclass correlation coefficient of 0.99). Nonparametric tests were used to assess differences (Mann-Whitney U-Test and Hodges Lehmann estimator); confidence intervals for percentages were calculated by inverting the score test. All tests were two-sided and used an alpha of p=0.05 as the threshold for significance. All analyses were conducted in R 3.4.4. The study was exempted from review by the relevant institutional IRB

Results: 37 of 124 articles [29.8% (CI: 22.1-38.8%)] reported both positive and negative likelihood ratios, in contrast to 121 of 124 which reported sensitivity and specificity. Although more papers report sensitivity and specificity than likelihood ratios, for papers that do report each statistic, approximately the same percentage report confidence intervals: 84 articles [69.4% (CI: 60.3-77.3%)] of 121 for sensitivity/specificity, and 27 articles [73% (CI: 55.6-85.6%)] of 37 for likelihood ratios

Conclusion: This analysis provides systematic evidence of diagnostic test reporting in the EM literature. More than 1 in 3 articles fails to report confidence intervals around sensitivity and specificity. LRs and their associated 95% CI are infrequently reported. Improved reporting can enhance the usage of diagnostic tests in emergency medicine.

414 Not Enough Shades of Gray: Shortcomings in Systematic Reviews of Emergency Medicine Literature

Zachary Pope[1], Rohan Datta[1], Austin J. Parish[1], and Jason R. West[2]
[1]*Lincoln Medical and Mental Health Center,* [2]*Lincoln Medical and Mental Health Center, Weill Cornell Medicine*

Background and Objectives: The validity of systematic reviews (SRs) depends deeply on the search strategies they employ, among other things. We sought to determine the extent to which recent SRs in emergency medicine 1) search the gray literature, 2) search clinical trial registries, 3) utilize manual searches of reference lists for additional studies, 4) discuss publication bias and 5) follow standard SR reporting guidelines

Methods: We searched PubMed for SRs published between January 1st 2009 and April 30th 2019 across the top 5 EM journals by H-index, including publications published online ahead of print. All manuscripts were examined by the authors to confirm that they were SRs. For each included SR, the following data was extracted: year of publication, whether the SR focused on a therapeutic intervention or diagnostic test, whether gray literature was searched, whether a hand search was employed, whether publication bias was mentioned, whether a clinical trial registry was searched, and whether a reporting guideline was followed. We defined a "complete" search strategy as searching grey literature, using a hand search, and searching trial registries

Results: We identified 578 articles for inclusion, of which 29 were excluded because they were determined to not be SRs. Of the 549 included SRs, 307 (55.9%) were therapeutic, 130 (23.7%) were diagnostic, and 112 (20.4%) were other SR types such as studies of risk factors. Overall, of the 549 SRs, 206 (37.5%) searched the grey literature, 427 (77.8%) utilized a hand search, 211 (38.4%) searched a clinical trial registry, 297 (54.1%) mentioned publication bias, and 351 (63.9%) utilized reporting guidelines. Although the odds of mentioning reporting guidelines ($p < 0.0001$) and publication bias ($p = 0.028$) increased