Comparative Effectiveness Review Disposition of Comments Report

Title: *Diagnostic Errors in the Emergency Department: A Systematic Review*

Draft report available for public comment from March 1, 2022, to March 29, 2022.

Citation: Newman-Toker DE, Peterson SM, Badihian S, Hassoon A, Nassery N, Parizadeh D, Wilson LM, Jia Y, Omron R, Tharmarajah S, Guerin L, Bastani P, Fracica ES, Kotwal S, Robinson KA. Diagnostic Errors in the Emergency Department: A Systematic Review. Comparative Effectiveness Review No. 258. (Prepared by the Johns Hopkins University Evidence-based Practice Center under Contract No. 75Q80120D00003.) Rockville, MD: Agency for Healthcare Research and Quality. December 2022. Errata and Addendum, August 2023. DOI: 10.23970/AHRQEPCCER258. Posted final reports are located on the Effective Health Care Program search page.

Comments to Draft Report

The Effective Health Care (EHC) Program encourages the public to participate in the development of its research projects. Each draft report is posted to the EHC Program website or AHRQ website for public comment for a 3- to 4-week period. Comments can be submitted via the website, mail, or email. At the conclusion of the public comment period, authors use the commentators' comments to revise the draft report.

Comments on draft reports and the authors' responses to the comments are posted for public viewing on the website approximately 3 months after the final report is published. Comments are not edited for spelling, grammar, or other content errors. Each comment is listed with the name and affiliation of the commentator, if this information is provided. Commentators are not required to provide their names or affiliations in order to submit suggestions or comments.

This document includes the responses by the authors of the report to comments that were submitted for this draft report. The responses to comments in this disposition report are those of the authors, who are responsible for its contents, and do not necessarily represent the views of the Agency for Healthcare Research and Quality.

Summary of Peer Reviewer Comments and Author Response

This research review underwent peer review before the draft report was posted for public comment on the EHC website. We received comments from eight technical experts and four peer reviewers. Below is a summary of the more substantive edits we made based on the peer review comments.

- We provided additional context for the overall diagnostic error rates in any medical setting.
- We clarified how the conditions of interest used for some of the questions were specified *a priori*.
- We provided definitions and/or clarifications for the terms used in the report (e.g., diagnostic error, misdiagnosis-related harms). Additionally, we clarified that diagnostic errors refers to both false negatives and false positives.
- We clarified how we calculated our estimates of diagnostic errors and serious misdiagnosis-related harms.
- We added more context about diagnostic errors in children.
- We provided more discussion about the strengths and weaknesses of the studies that were included for KQ1 ("What clinical conditions are associated with the greatest number and highest risk of emergency department (ED) diagnostic errors and associated harms?") and included uncertainty estimates.
- We provided additional context for preventable diagnostic errors.
- We added teamwork as a factor that could influence diagnostic errors.
- We discuss the lack of evidence on how patients can influence diagnostic errors.
- We added more text about the applicability of the results.
- We added more discussion about unintentional consequences that could arise when attempting to address diagnostic errors.
- We added some text on approaches to measuring diagnostic errors at the institutional level.
- We provided more context about how a dashboard could impact diagnostic error rates.

We developed a Frequently Asked Questions (FAQ) document to address common questions raised by the peer and public review (Newman-Toker DE, Peterson SM, Badihian S, et al. Frequently Asked Questions for "Diagnostic Errors in the Emergency Department: A Systematic Review." Open Science Framework. 2023. DOI: https://doi.org/10.17605/OSF.IO/B7XVM).

Source: https://effectivehealthcare.ahrq.gov/products/diagnostic-errors-emergency/research
Published Online: December 15, 2022, Errata and Addendum August 14, 2023

Public Comments and Author Response

Commentator & Affiliation	Section	Comment	Response
Jonathan Edlow	General comment on Stroke misdiagnoses	So with respect to stroke, while it is clear that missed strokes are usually less severe and apparent on initial presentation, and usually more severe and more obvious on representation, the notion that correctly diagnosed strokes will end up without a deficit is clearly not true. IV thrombolysis will result in an ~ 12% absolute risk reduction of a significant neurological deficits (mRS = 0-1) so the majority of correct diagnosed and treated stroke patient still end up with whatever deficit they were going to have. This is less true with endovascular treatments for LVO occasion strokes (the treatment effect is more powerful) but as you note, they are less likely to be misdiagnosed.	Early diagnosis generally leads to early secondary prevention (e.g., dual antiplatelet therapy), not acute treatment (e.g., IV thrombolytics or endovascular therapy). Secondary prevention to avoid major stroke after minor stroke/TIA is far more effective than acute treatment for major stroke. We have now clarified this point in the Discussion ("Instead, it is essential to create mechanisms that rapidly identify patients with subtle stroke symptoms which are prone to be missed (e.g., dizziness and headaches), in order to bring such patients into the stroke treatment pathways so they too may benefit from prompt therapy (e.g., dual antiplatelet therapy for early secondary prevention, which, if applied in the first 24 hours, lowers risk of major stroke after minor stroke or TIA by 34% over the next 21 days)")

Commentator & Affiliation	Section	Comment	Response
Jonathan Edlow	General comment	You partly address this but it's important to ask "Who is calling it an error?" (An internist reviewing an ED chart vs an emergency physician) and WHAT is the gold standard for the diagnosis? (You address this too when you mention that in ?4% of ED admits to the hospital, it was the original ED diagnosis that was right and the hospital diagnosis was wrong). I don't mean this comment to suggest "partisanship" for emergency physicians but just to state the obvious complexity of it all. Our department has been studying use of a rule-based system for defining error. That is to so, if you can't state a rule that was broken, it's not an error (even if there was an adverse event). Examples, "ALL women of childbearing years with abdominal pain should have a UCG" or "ALL patients with chest pain over the age of 40 should have an ECG", etc. In our QA committee we explicitly pose the question, can we articulate a rule that was broken. Admittedly it gets harder the more subtle the case and if we cannot articulate a specific rule, we call in a judgment call (with or without an adverse event). It helps to reduce the hindsight bias.	This comment addresses the subjectivity inherent in determining whether there was a process failure during the diagnostic process. It is for this reason that we used the National Academy of Medicine's definition of diagnostic error. The NAM definition of diagnostic error does not require a process failure for a mistaken diagnostic label to be considered a diagnostic error. This reduces subjectivity in the determination of diagnostic errors. We have added a section to the Limitations to address this issue (and related measurement concerns expressed by another public commenter): "Most studies did not directly address issues surrounding measurement of diagnostic error (e.g., validity, reliability, determination of causes, preventability, or attribution of harms). In clinical practice, many disease reference standards are insufficiently understood, developed, and implemented, so diagnosticians often disagree on final patient diagnoses. To the extent that manual chart reviews were used to identify errors, original studies are likely to suffer from problems of poor chart documentation, low inter-rater reliability, and hindsight bias. The problem of author bias in choice of definition or method of measurement (e.g., specialists [or diagnostic error "advocates"] determining ED misdiagnosis and favoring more lax definitions of error/harm, or the reverse, with ED clinicians favoring more stringent definitions) is difficult to ascertain. Our use of the NAM definition of diagnostic error mitigates some of these concerns, since there is less subjectivity inherent in a diagnostic label change (e.g., discharged with "musculoskeletal chest pain" returns with "aortic dissection" within 24 hours) than in the determination of preventability, which is known to be highly subjective. Also, many included studies used stringent measurement protocols or objective statistical methods (e.g., SPADE). Nevertheless, poorly standardized or low-reliability measurements are important limitations."

Commentator & Affiliation	Section	Comment	Response
Jonathan Edlow	General comment	I think that MI misdiagnosis has fallen because we have ECGs and troponin AND an acute treatment. Ditto with ectopic pregnancy which interestingly is not on the top 10 list (although I suspect that 40 years ago it may have been) because now, we have UCG and US.	We have added specific mention of electrocardiograms and troponins to the sentence in the Introduction that already addressed this issue "Diagnostic error rates for myocardial infarction, for example, are impressively low at about 1 to 2 percent12 (in part due to the availability of electrocardiograms and a reliable lab test [i.e., troponin assays])."
Jonathan Edlow	General comment	The dizziness quandary is a special interest of mine. You may be interested to know that we are about half way though the 14-16 month process of completing the SAEM GRACE-3 project on acute dizziness which will feature a cool multi-media educational module. Re: SAH, the Waxman article not-withstanding, I think that the preponderance of indirect evidence shows we are missing it LESS frequently than more if one takes a longer time horizon from the late 1980s and again, this is due to more CT scanning. As for dizziness, there is just a huge knowledge gap.	Thank you for sharing this information. We mentioned the SAEM guideline development process in the Discussion "Improving diagnosis of strokes in dizziness is a top priority for ED clinicians, and a clinical practice guideline for acute dizziness diagnosis is currently under development by the Society for Academic Emergency Medicine. " Addressing trends since the late 1980s is outside of the scope of our review. We included studies from 2000 to present.

Commentator & Affiliation	Section	Comment	Response
Jonathan Edlow	General comment on spinal epidural abscess	And regarding spinal epidural abscess, to some extent, psychologically (not necessarily at a conscious level), there is a reluctance to 'order a test' that is a pain in the ass to get (either begging the radiologist, or having to wait for hours & hours or having to transfer a patient off-hours to get the study at another facility. So there are interrelationships between “not ordering a test” and “not having the test EASILY available”. Ditto with consultants (by the way, the Royl study on dizziness was in an ED but all patients were consulted on by neurologists so I am not sure it can be counted as ED misses - in fact, I think it means that dizziness is just complicated and that the knowledge gap is shared by many specialties and includes a horrible diagnostic algorithm that dates back to 1972. So these factors (off-hours, sub-optimal availability of consultant or tests (MRI), even if they are theoretically available interweave with one another and I believe play a role in misdiagnoses. It underscores the reality that an ED operates 24x7 whereas the rest of the hospital, to a variable but large extent, operates 9-5 M-F.	We acknowledge the challenges and complexity of providing care in the ED. The report is about ED care, not ED physician diagnostic performance, so whatever happens (e.g., consultations) or does not (e.g., lack of consultants or tests) in the ED is part of that ED care, so “counts.” The paragraph in the Introduction on reasons why ED errors occur acknowledges this issue of test unavailability ("There are many reasons why the ED may be the most challenging clinical setting for diagnosis. ... Many EDs have limited access to specialty consultants or advanced diagnostic tests, such as magnetic resonance imaging."). Access to consultation and testing are also defined in Table 13 as prospective predictors of diagnostic error and described in the response to KQ3 and Research Recommendations ("For KQ3, more research needs to be done to clarify the extent to which structural factors (particularly those that could be induced to change by payment mechanisms) are strong predictors of diagnostic error and harms. For example, these might include ED discharge fraction, staffing patterns (e.g., volumes per clinician, routine availability of consultants), and access to specialized imaging or diagnostic laboratory tests.").

Commentator & Affiliation	Section	Comment	Response
Jonathan Edlow	Comment on the 72-hour return metric	Last, the 72-hour return metric is so tricky because it includes patients who were recommended admission but refused, patients who came back for unrelated problems, patients who came back because Plan A wasn't working (even though it was a reasonable plan), patients who were admitted to the hospital then returned to ED all within 3 days, patients whose misdiagnoses were not gettable on Visit #1 and then misdiagnoses that SHOULD HAVE been picked up on visit #1"	As described in Figure 14, only a subset of 72-hour returns are related to diagnostic errors.
Charles A Pilcher	List of top 10 diseases	I would add 'necrotizing fasciitis' to the list of the top 10, at least in terms of morbidity and mortality if missed.	We appreciate that many conditions might be added to the list of considerations (e.g., obstructive hydrocephalus, compartment syndrome), among them necrotizing fasciitis. However, this particular one was not identified as part of our preliminary search for the top harmful conditions, nor by the Technical Expert Panel or Key Informants who provided input on the review scope and methods. We have added text to the Strengths and Limitations section on this point ("However, because of the constrained focus on the most common conditions, we do not have data on misdiagnosis of less common conditions that may nevertheless be of importance to ED clinicians (non-accidental trauma, necrotizing fasciitis, compartment syndrome, brain tumors, obstructive hydrocephalus, ovarian torsion, post-partum hemorrhage, etc.); this is a limitation."). We have also included necrotizing fasciitis as part of a footnote to Table 3 (it represented 1.2% of all serious misdiagnosis-related harms in malpractice claims, where it was the 18th ranked condition causing serious misdiagnosis-related harms) and as a line item in Table 4.

Commentator & Affiliation	Section	Comment	Response
Joseph A. Grubenhoff	Methods/top 10 diseases	As a pediatric EM academic physician, I agree with the statement that cardiac disease and testicular (and OVARIAN which is omitted) torsion are a highly relevant conditions for diagnostic delay in the ED. However, I am surprised that the authors include necrotizing enterocolitis (NEC).NEC is almost exclusively a disease of very premature neonates. In 17 years working in the pediatric ED of a top-five children's hospital I have never seen a case of NEC. While anecdote is not evidence and term infants are at risk of NEC, it is not a common pediatric ED problem.	We appreciate that necrotizing enterocolitis is almost exclusively a disease of premature neonates. We also recognize this is principally a problem encountered in the NICU, rather than the ED. However, it has been reported in the ED, now that pre-term infants are being discharged from the hospital (PMID: 11489407). NEC was not on our original list derived from the preliminary literature search, but it was identified as part of our discussions with, and feedback from, the Technical Expert Panel and Key Informants, so it was included. Ovarian torsion was also discussed with the Technical Expert Panel and Key Informants, but available evidence prior to the review (from malpractice claims) listed testicular torsion and not ovarian torsion. As noted in the Strengths/Limitations section, "Overall, there is a relative paucity of literature on diagnostic errors among pediatric ED populations. More studies are warranted, including research on how the distribution of diseases (KQ1), rates of diagnostic error (KQ2), and causes/risk factors (KQ3) differ from those in adult patients." To strengthen this point and elaborate on ovarian torsion, we have added the following language to the Gap analysis for KQ1: "Some diseases relevant to children were not identified in our preliminary search or through our Technical Expert Panel and Key Informant interview processes, so were not explicitly assessed in our protocol (e.g., ovarian torsion, child abuse, brain tumors); these may be important to future inquiries."

Commentator & Affiliation	Section	Comment	Response
Joseph A. Grubenhoff	Methods/top 10 diseases	On the other hand, the authors do not include NON-ACCIDENTAL TRAUMA/CHILD ABUSE. There is ample literature indicating that child presenting to the ED with severe injuries typically have presented previously with minor (sentinel) injuries that, were they recognized and acted upon during the prior encounter, may have led to protection of the child. Citations below. PLEASE STRONGLY CONSIDER ADDING CHILD ABUSE TO YOUR STATEMENT (...) See: 1. Thorpe EL, Zuckerbraun NS, Wolford JE, Berger RP. Missed opportunities to diagnose child physical abuse. Pediatr Emerg Care. 2014 Nov;30(11):771-6. doi: 10.1097/PEC.0000000000000257. PMID: 25343739. 2. Sheets LK, Leach ME, Koszewski IJ, Lessmeier AM, Nugent M, Simpson P. Sentinel injuries in infants evaluated for child physical abuse. Pediatrics. 2013 Apr;131(4):701-7. doi: 10.1542/peds.2012-2780. Epub 2013 Mar 11. PMID: 23478861. 3. Lindberg DM, Beaty B, Juarez-Colunga E, Wood JN, Runyan DK. Testing for Abuse in Children With Sentinel Injuries. Pediatrics. 2015 Nov;136(5):831-8. doi: 10.1542/peds.2015-1487. Epub 2015 Oct 5. PMID: 26438705.	Child abuse was not on our original list derived from the preliminary literature search, nor was it identified as part of our Technical Expert Panel and Key Informant input. As noted in the Strengths/Limitations section, "Overall, there is a relative paucity of literature on diagnostic errors among pediatric ED populations. More studies are warranted, including research on how the distribution of diseases (KQ1), rates of diagnostic error (KQ2), and causes/risk factors (KQ3) differ from those in adult patients." To strengthen this point and elaborate on child abuse, we have added the following language to the Gap analysis for KQ1: "Some diseases relevant to children were not identified in our preliminary search or through our Technical Expert Panel and Key Informant feedback processes, so were not explicitly assessed in our protocol (e.g., ovarian torsion, child abuse, brain tumors); these may be important to future inquiries." We have cited the suggested references in this new sentence.

Commentator & Affiliation	Section	Comment	Response
Joseph A. Grubenhoff (cont'd)	Methods/top 10 diseases (cont'd)	Letson MM, Cooper JN, Deans KJ, Scribano PV, Makoroff KL, Feldman KW, Berger RP. Prior opportunities to identify abuse in children with abusive head trauma. Child Abuse Negl. 2016 Oct;60:36-45. doi: 10.1016/j.chiabu.2016.09.001. Epub 2016 Sep 25. PMID: 27680755.	(response above)
Joseph A. Grubenhoff	Methods/top 10 diseases	Also, BRAIN TUMORS are the MOST COMMON CHILDHOOD CANCER but typically present with vague and common symptoms (especially headache and vomiting) that most of the time do not herald serious disease. There is a multicenter Pediatric Emergency Care Applied Research Network study on-going to identify very low risk criteria for serious intracranial diseases similar to the seminal PECARN work to identify children at very low risk of clinically important TBI. Given the amount of funding/effort and prioritization of identifying low risk patients presenting with HA, would advise including brain tumors in the statement.	Brain tumor was not on our original list derived from the preliminary literature search, nor was it identified as part of our discussions with the Technical Expert Panel and Key Informants. As noted in the Strengths/Limitations section, "Overall, there is a relative paucity of literature on diagnostic errors among pediatric ED populations. More studies are warranted, including research on how the distribution of diseases (KQ1), rates of diagnostic error (KQ2), and causes/risk factors (KQ3) differ from those in adult patients." To strengthen this point and elaborate on brain tumor, we have added the following language to the Gap analysis for KQ1: "Some diseases relevant to children were not identified in our preliminary search or through our Technical Expert Panel and Key Informant interview processes, so were not explicitly assessed in our protocol (e.g., ovarian torsion, child abuse, brain tumors); these may be important to future inquiries." We have also added the following sentence to a new paragraph in the Strengths and Limitations section related to the list of diseases... "However, because of the constrained focus on the most common conditions, we do not have data on misdiagnosis of less common conditions that may nevertheless be of importance to ED clinicians (non-accidental trauma, necrotizing fasciitis, compartment syndrome, brain tumors, obstructive hydrocephalus, ovarian torsion, post-partum hemorrhage, etc.); this is a limitation." We have also now included brain and spinal tumors as part of a footnote to Table 3 (representing 1.4% of all serious misdiagnosis-related harms in malpractice claims, where it was the 16th ranked condition causing serious misdiagnosis-related harms).

Commentator & Affiliation	Section	Comment	Response
Joseph A. Grubenhoff	Method/Top 10 diseases (using malpractice lawsuits)	Another major flaw in using a malpractice claims to identify conditions at greatest risk of diagnostic error among pediatric patients is the fact that children typically have fewer comorbidities complicating their recovery (heart disease, hypertension, obesity, etc). Thus, they are more likely to recover and less likely to sue because they don't suffer PERMANENT harm. But many children have conditions that, when missed, lead to longer hospitalizations, more serious disease (e.g. missed osteomyelitis that leads to bacteremia, sepsis and multifocal osteo), and invasive procedures that they eventually recover from.	We describe in some detail in the report the nature of bias in malpractice claims in the section entitled "Representativeness of Malpractice Claims Data for Disease Distribution," which are clearly biased towards more severe harms ("In particular, claims are known to be biased towards higher-severity harms; this is self-evident from Tables 3 and 4, since high-severity harms are relatively rare, yet among the malpractice cases there are more high-severity harm cases than low- and medium-severity harm cases combined. This is further reinforced by the much higher fraction of high-severity harms in the malpractice claims than in the large incident report study described above (58% versus 15%).") As we have mentioned in several locations in the report as to limitations, the report disproportionately reflects serious harms, rather than less serious harms (Strengths and Limitations: "On KQ1 (diseases), the literature was relatively strong for diseases causing more severe harms but fairly weak on the disease distribution for lower-severity errors."). We address the potential age bias in use of malpractice claims in the report in the Discussion:- "If the principal mechanism by which lung cancer is missed in the ED is via missed incidental lung nodules on chest X-ray, then there is no specific reason why this should occur with greater frequency in younger patients than older ones—if anything, they should have less lung pathology that interferes with radiographic interpretation. This suggests a likely age bias to file a legal claim when the patient is younger, rather than older."). We have added language to the section of KQ1 devoted to "Difference by Patient Age Group" about the possible greater resilience of pediatric patients ("...(b) harms are less frequent among children (e.g., because they are less often impacted by life-threatening diseases or are more medically resilient when such diseases are present).").

Commentator & Affiliation	Section	Comment	Response
Joseph A. Grubenhoff	Using Malpractice Suits/Results	Another problem with the use of large malpractice claims datasets is the skew toward adult cases. There are many more adults in the US than children. Children of color and economically disadvantaged children often rely heavily on the ED for primary care and who cannot afford to bring suits (or indeed wouldn't have the knowhow to even start). So, the top ten list is inherently skewed leaving out a very vulnerable population.	As we describe in some detail in the section in KQ1 "Differences by Patient Age Group" there is no reason to believe that malpractice claims are skewed towards adults (i.e., overrepresent adults) or that serious harms are overlooked in children. We concur that there are more claims among adults because there are more adult ED patients and adults more often have dangerous diseases causing their symptoms; as noted by the commenter, other factors may be at play (e.g., likelihood of permanent harm to a child being lower, given a misdiagnosis). We acknowledge that the ED is often utilized by people of color and the economically disadvantaged (whether children or adults), and these vulnerable populations (i.e., minorities/low SES/low health literacy) may be underrepresented in claims. We have added language to the section "Representativeness of Malpractice Claims Data for Disease Distribution" on this point ("Other biases could be at work that are not readily apparent from the available literature. For example, disadvantaged or vulnerable populations (e.g., those who are differently abled, racial or ethnic minorities, lower health literacy, lower socioeconomic status, prisoners) might be both more likely to be misdiagnosed and less likely to file a legal claim. However, we could find no specific evidence to suggest that this would likely impact the distribution of diseases for KQ1.").

Commentator & Affiliation	Section	Comment	Response
Joseph A. Grubenhoff	Using Malpractice Suits/Results	On pages 21-22, the "results" regarding low proportions of pediatric misdiagnosis malpractice cases include a fair amount of speculation. Indeed, a few sentences related to the severity of injury draw a fairly drastic assumption that families would be more likely to seek reparations due to "devastating medical misdiagnosis" and not being able to "live a full life". My experience as a pediatric EM physician does not bear that out. Many of the children seeking emergency care are children of color and economically disadvantaged. As such, they may not have the financial means and necessary knowledge to even begin a suit.	Thank you for raising this concern. We added language to the section "Representativeness of Malpractice Claims Data for Disease Distribution" on this point ("Other biases could be at work that are not readily apparent from the available literature. For example, disadvantaged or vulnerable populations (e.g., those who are differently abled, racial or ethnic minorities, lower health literacy, lower socioeconomic status, prisoners, immigrants) might be more likely to be misdiagnosed and less likely to file a legal claim. However, we could find no specific evidence to suggest that this would likely impact the distribution of diseases for KQ1. In particular, it is important to note that there was almost complete alignment between the list of diseases from malpractice claims and those reported in diagnostic safety incidents (Table 2), which argues fairly powerfully against a major disease maldistribution based on claims data."). As to the issue of "speculation" on the reason why claims are less frequent among children than adults, we have bolstered the passage below with additional supporting citations in the body of the report (*citations not shown here*): "Although this absolute frequency difference between children and adults could be accounted for by a lower likelihood of a lawsuit being brought when the patient is a child, this seems highly improbable; if anything, one would suspect just the opposite, since legal actions are disproportionately sought when the severity of adverse outcomes is greater (as would be the case for a child who might otherwise have a “full life to live” were it not for a devastating medical misdiagnosis). The greater likelihood of a lawsuit being brought when the claimant is a child is supported by data from the National Practitioner Data Bank showing higher payouts in pediatric than adult cases, with the highest payouts occurring among the youngest children and the lowest payouts among the oldest adults. Some specific data on the relative frequency of claims, such as those related to lung cancer misdiagnosis in the ED, appear to confirm the general suspicion of a higher likelihood that cases will be brought when patients are younger (see Representativeness of Malpractice Claims Data for Disease Distribution, below)."

Commentator & Affiliation	Section	Comment	Response
Joseph A. Grubenhoff	Child abuse/results	As mentioned, there is also a major missing category of pediatric disease that will likely never result in malpractice, that of CHILD ABUSE. The most common abusers are household contacts (parents, paramours, second degree relatives) and there is almost always going to be missing information (the history in children who have been abused is inherently limited because an abuser is rarely going to offer that info due to criminal liability). Add to that the fact that the prior literature on missing sentinel injuries, and the scenario is very UNLIKELY to result in a law suit. Simply stated, an abuser is not going to file a lawsuit for a doc missing child abuse because that runs the risk of being jailed if not already. These children sustain some of the most serious life long harms and death and will be almost entirely missed by methods relying on malpractice data.	We appreciate and concur with this critique. We have added specific language to this effect to the section about "Representativeness of Malpractice Claims Data for Disease Distribution" ("Child abuse (non-accidental trauma) is a special case in which misdiagnoses are unlikely to result in malpractice claims, even if the underlying problem does result in serious harms to the child, since the abuser (often a parent) is unlikely to draw attention to the condition via a legal claim."). We have also added a paragraph to the KQ1 Gaps section of the report that calls out child abuse and also identifies this class of problems (i.e., those not likely to be reported when missed) as a specific problem ("The special case of child abuse (which was not incorporated into our study design but was identified during the review/comment period) highlights an important gap around recognition of diagnostic errors for diseases that may be intentionally concealed, rather than surfaced, as problems. The Centers for Disease Control and Prevention have estimated that nearly 1 in 7 children suffer abuse and neglect, resulting in 1,750 deaths in the United States in 2020. One older study of 173 abused children younger under age 3 with head injuries found 54 (31%) were not recognized by physicians (across settings) as non-accidental injuries; among these, 15 (28%) were reinjured after the misdiagnosis. A more recent, multi-center, ED-based study in the Netherlands found that EDs complying with screening guidelines for child abuse were 4-fold more likely to detect cases (0.3% versus 0.1%, $p<0.001$), suggesting that many missed cases are likely detectable. Because abusive parents are highly unlikely to file a malpractice claim for an ED missed diagnosis of abuse, malpractice data will grossly underrepresent this condition. The same is likely to be true for other forms of abuse (e.g., missed spousal abuse, elder abuse), certain socially unacceptable conditions (e.g., missed cases of illicit drug use or dependence), or factitious disorders (e.g., missed Munchausen syndrome). Furthermore, individuals may be more likely to seek care at different EDs, limiting the utility of single institutions to detect missed cases (e.g., via chart review). For these disorders, special efforts must be made to identify misdiagnoses using alternative data sources and methods.").

Commentator & Affiliation	Section	Comment	Response
Joseph A. Grubenhoff	Results (general)	As this "Results Section" should be results, strongly recommend limiting the speculation and moving it to the Discussion.	We have revised the Results section to move some of the interpretation to Discussion but feel that the reader is helped by retaining some of the interpretation in the Results. We have added citations to support the interpretation, such as in the section called out by the commenter previously.
Joseph A. Grubenhoff	General Comments	After reading this report, I am left feeling that the pediatric ED population is not well understood by the authors. Were there any pediatric ED physicians involved in drafting this document? Because this will be an AHRQ release, it will clearly be used/referenced to set policy for the agency in terms of funding priorities. Children are at risk of diagnostic error but, one look at the DEM conference attendance and it's clear that pediatrics is not yet well represented in the research and QI space. Because so much emphasis is being placed on malpractice claims data, which the report admits is skewed, the priorities for investigation set by AHRQ and non-profit foundations that follow their lead (e.g. Moore Foundation) will continue to under-fund pediatric diagnostic error research. This report appears to be an expanded version of Newman-Toker's "Big Three" paper. That work seems to have taken on a life of its own in the diagnostic error space and is now driving larger and larger policy decisions that will lead to skewed research priorities for a few decades. Children, just like in other areas of research, are going to get left behind.	Two pediatric ED physicians were on the Technical Expert Panel informing the design and conduct of the systematic review. They both also reviewed the report as external reviewers and their critiques were addressed by the authorship team in earlier drafts. Malpractice claims only feature prominently in the response to KQ1 and not at all in KQ2; while malpractice claims are biased towards high-severity outcomes, there is little reason to believe that this leads to any of the important differences in the list of conditions between children and adults (which are almost certainly the result of differences in disease prevalence by age). The fact that there is less literature describing pediatric ED error than adult ED error is a gap confirmed by this systematic review, and one that we have called out specifically as such ("Overall, there is a relative paucity of literature on diagnostic errors among pediatric ED populations."). We appreciate the commenter's concern that diagnostic errors in pediatric EM differ from those in adult EM practice (including "other" non-Big Three diseases, which are, as shown in Figure 2 and Table 5, are more common among children as a proportion of all diagnostic errors). This point is made repeatedly throughout the report, and further research into pediatric diagnostic error research is expressly called for ("More research is also needed to better characterize the diseases associated with diagnostic error in pediatric ED settings and specialty EDs, where there are many fewer studies.").

Commentator & Affiliation	Section	Comment	Response
Anonymous	General Comments	This systematic review has several limitations that make its main findings and implications invalid and highly biased. Some of the methodological “fatal flaws” include 1) handpicking areas and diseases to focus on prior to conducting the review rather than relying on the literature, 2) using largely “numerator-only” data to make most assessments and answering key questions, 3) cherry-picking selected studies while ignoring others to make estimates, and 4) relying on studies that did not study diagnostic error to make key estimates. All of these issues raise substantial scientific concerns for an AHRQ sponsored review, findings of which are currently either questionable, meaningless or both. In addition, the implications and policy section is largely disconnected from the systematic review and seem to be focused on furthering a highly selective and narrow personalized agenda.	Thank you for sharing your concerns. (1) "Handpicking areas and diseases to focus on": As described in our protocol and the Methods section of the report, the list of diseases under consideration was formulated *a priori* on the basis of prior literature and informed by input from a Technical Expert Panel and Key Informants. Now having analyzed the results from KQ1, the prespecified list appears to have been fairly complete vis-a-vis the most common causes of misdiagnosis-related harms - for example, in the largest incident report study of ED diagnostic errors (n=2,288) (which was not used to determine the prespecified list), all top 12 conditions found in that study (see Table 1 from PMID: 31801474) appeared in our prespecified list. No other conditions identified in that study had higher individual frequency, and, collectively, those "other" conditions outside the top 12 accounted for just 30% of the total incidents reported (n=679/2,288). While some conditions (particularly those affecting children) may have been underrepresented (e.g., missed child abuse/non-accidental trauma), we found no evidence to suggest that using a prespecified list based on prior literature and input from the Technical Expert Panel and Key Informants appreciably affected the overall results. We have added a paragraph to the Limitations section describing this issue.

Commentator & Affiliation	Section	Comment	Response
Anonymous (cont'd)	General Comments (cont'd)	(comment above)	(2) Use of "numerator only" data: Malpractice claims and incident reports (i.e., two commonly used forms of "numerator only" data) cannot and were not used to answer rate questions (i.e., KQ2) --- all rates were from numerator-denominator studies. Having a denominator is not methodologically relevant for answering KQ1 (list of diseases) - the key methodological issue for KQ1 is whether the source data are biased with respect to the list of diseases reported, not whether the precise denominator (i.e., source population) is known. Sources of bias in malpractice claims data are discussed under KQ1 ("Representativeness of Malpractice Claims Data for Disease Distribution"). The claims-based list is corroborated independently by the presence of the same list of diseases from incident report data, as mentioned in response to part (1) of this comment. It is an interesting question whether root causes (KQ3) identified in malpractice claims might be biased. We have added some text on this issue ("Representativeness of Malpractice Claims Data for Root Causes. It is known that malpractice claims data represent a biased sample of cases, so it is then reasonable to consider whether bias(es) might influence the root causes of diagnostic error identified. As described above, it was clear from ED incident report studies (e.g., Okafor, 2016; Hussain, 2019) that the spectrum of root causes identified is quite similar to that found in ED malpractice claims studies—mostly cognitive errors related to bedside diagnostic decision-making (especially clinical examination, test ordering, or integration of test results into diagnostic reasoning). What is not known is whether both malpractice claims and voluntary incident reports might be biased towards cases with cognitive errors by physicians. This question cannot be easily addressed by retrospective studies relying on chart review, since most potential root causes must be inferred (i.e., they are not actually captured or recorded). Nor can it be addressed by diagnostically oriented, experimental vignette-based studies (which only assess for cognitive errors). To address this question rigorously, one would need a cohort study or clinical trial that prospectively captured all potential root causes and then assessed diagnostic errors and root causes. We found no such studies, so this remains an unanswered scientific question.").

Commentator & Affiliation	Section	Comment	Response
Anonymous (cont'd)	General Comments (cont'd)	(comment above)	(3) "cherry picking" studies: As in all rigorous systematic reviews, the research methods, inclusion/exclusion criteria, and plans for synthesis were described *a priori* in the protocol. Part of the analytic work is to identify methodological heterogeneity and to determine if studies should be synthesized in meta-analytic fashion (as noted in the protocol, "We conducted meta-analyses when there were sufficient data (i.e., at least two studies) and studies were sufficiently homogenous with respect to key variables (e.g., population characteristics, condition, provider type, and data source/study design)."). It is methodologically incorrect to combine estimates from studies that are meaningfully heterogeneous in underlying methods, so some studies that used different designs were not meta-analytically synthesized (e.g., retrospective studies sampling only "return visit" cases were not synthesized with prospective studies of consecutive, unselected patients). (4) "relying on studies that did not study diagnostic error": All studies captured by the systematic review met the National Academy of Medicine definition of diagnostic error, as defined in the Methods section of the report. (5) "implications and policy disconnected from the systematic review" - the offered considerations for policy derive directly from the review findings: "(1) standardizing measurement and research results reporting to maximize comparability of measures of diagnostic error and misdiagnosis-related harms" - this derives directly from the lack of standardized measurement of diagnostic error and harms identified by the systematic review; "(2) creating a National Diagnostic Performance Dashboard to track performance (analogous to the Dartmouth Atlas Project for utilization of healthcare services)" - this derives from the lack of adequate national benchmarking and lack of comparability of measurement across EDs identified in this systematic review; and "(3) using multiple policy levers (e.g., research funding, public accountability, payment reforms) to push for the rapid development and deployment of solutions that address this major patient safety and quality problem" - this derives directly from the overall public health scale/scope of the problem identified by the review.

Commentator & Affiliation	Section	Comment	Response
Anonymous	Introduction	Introduction should better frame the complexity of defining and measuring diagnostic error	Thank you. There is discussion of the complexity of defining and measuring diagnostic error in the Introduction section, and we have now juxtaposed the two paragraphs that deal with definitions and measurement issues to address the issue more forcefully ("Despite their toll on patients and society, diagnostic errors remain largely invisible. This is mostly because diagnostic errors are rarely evident at the time when they occur and only surface later, often when they are discovered by another clinician or after misdiagnosis-related harms have occurred. Furthermore, diagnostic errors are variably defined, difficult to measure, and not routinely tracked as part of patient safety or quality improvement initiatives. The National Academy of Medicine (NAM) defines diagnostic error as “the failure to (a) establish an accurate and timely explanation of the patient’s health problem(s) or (b) communicate that explanation to the patient.” Notably, this definition (which is used in this report) does not require a care process failure (e.g., a specific clinical reasoning “mistake” on the part of an individual clinician) and is agnostic with respect to any resulting harms or their preventability. Furthermore, it does not elaborate on the words “accurate” or “timely,” nor does it draw distinctions between false negative and false positive errors or specify how management differences might be used inferentially in assessing the “correctness” of diagnostic decision-making. There is no clear consensus on how to define “diagnostic error” at this deeper level, but some ED authors have made important attempts to do so. For example, a Swiss group examining diagnostic errors among admitted ED patients divided differences between ED and final hospital discharge diagnoses into those that were deemed, in their view, not to represent ED diagnostic errors (ED diagnosis was somewhat underspecified or a complication not present at the time of the ED visit became the primary inpatient diagnosis) and those that were considered diagnostic errors (ED missed a second, more important diagnosis or ED diagnosis was qualitatively incorrect).There is even less certainty about how best to capture communication failures between ED clinician and patient, and very few studies have sought to address this aspect of diagnostic error definitions. Whenever possible, we relied on the NAM definition of diagnostic error (e.g., to differentiate diagnostic errors from diagnostic errors with

Commentator & Affiliation	Section	Comment	Response
Anonymous (cont'd)	Introduction (cont'd)	(comment above)	adverse events or harms), but we also relied, as necessary, on individual study-based operational definitions, including more granular determinations of error, harms, and preventable harms that were used in the included studies.") We have also added a paragraph to the Limitations section on measurement concerns ("Most studies did not directly address issues surrounding measurement of diagnostic error (e.g., validity, reliability, determination of causes, preventability, or attribution of harms). In clinical practice, many disease reference standards are insufficiently understood, developed, and implemented, so diagnosticians often disagree on final patient diagnoses. To the extent that manual chart reviews were used to identify errors, original studies are likely to suffer from problems of poor chart documentation, low inter-rater reliability, and hindsight bias. The problem of author bias in choice of definition or method of measurement (e.g., specialists [or diagnostic error "advocates"] determining ED misdiagnosis and favoring more lax definitions of error/harm, or the reverse, with ED clinicians favoring more stringent definitions) is difficult to ascertain. Our use of the NAM definition of diagnostic error mitigates some of these concerns, since there is less subjectivity inherent in a diagnostic label change (e.g., discharged with "musculoskeletal chest pain" returns with "aortic dissection" within 24 hours) than in the determination of preventability, which is known to be highly subjective. Also, many included studies used stringent measurement protocols or objective statistical methods (e.g., SPADE). Nevertheless, poorly standardized or low-reliability measurements are important limitations.").

Commentator & Affiliation	Section	Comment	Response
Anonymous	Methods	On page 9, the authors say they decided to focus on selected conditions a priori most of which already fall into their “Big 3” categories that conveniently emerge later. On what basis did the authors exclude hundreds of possible studies that could have been included? Sounds like a sure shot way to bias a systematic review. This is a substantial problem and a fatal flaw of the review.	As described in the Methods, the list of diseases under consideration was formulated on the basis of prior literature and informed by input from a Technical Expert Panel and Key Informants. Now having analyzed the results from KQ1, the prespecified list appears to have been fairly complete vis-a-vis the most common causes of misdiagnosis-related harms - for example, in the largest incident report study of ED diagnostic errors (n=2,288) (which was not used to determine the prespecified list), all top 12 conditions found in that study (see Table 1 from PMID: 31801474) appeared in our prespecified list. No other conditions identified in that study had higher individual frequency, and, collectively, those "other" conditions outside the top 12 accounted for just 30% of the total incidents reported (n=679/2,288). While some conditions (particularly those affecting children) may have been underrepresented (e.g., missed child abuse/non-accidental trauma), we found no evidence to suggest that using a prespecified list based on prior literature and input from our Technical Expert Panel and Key Informants appreciably affected the overall results. We have added a paragraph to the Limitations section describing this issue.

Commentator & Affiliation	Section	Comment	Response
Anonymous	Methods	The review explores 'What are the most common and significant medical diagnostic failures in the ED, and why do they happen?' but relies heavily on 'numerator only' sources. In fact about half of the review's data sources are malpractice claims and incident reports, both of which cannot be used to estimate reliable frequencies/prevalence and are not representative of population level data. I suggest authors review Harvard Medical Practice, Utah Colorado studies as well as other rigorously done prevalence estimates in the patient safety literature. On page 7 the authors first suggest that 'numerator-only' labelling implies 'no explicitly defined source population from which they were drawn, so valid error/harm rates cannot be calculated'. Then claim in same para 'For KQ1, disease-agnostic data sources are needed, but numerator-only data are sufficient'. These are the same sources that inappropriately inform other key questions. Reliance on these sources raises significant concern and violates methodological knowledge that exists about measurement.	Malpractice claims and incident reports (i.e., two commonly used forms of "numerator only" data) cannot and they were not used to answer rate questions (i.e., KQ2) --- all rates were from numerator-denominator studies. Having a denominator is not methodologically relevant for answering KQ1 (list of diseases) - the key methodological issue for KQ1 is whether the source data are biased with respect to the list of diseases reported, not whether the precise denominator (i.e., source population) is known. Sources of bias in malpractice claims data are discussed extensively under KQ1 ("Representativeness of Malpractice Claims Data for Disease Distribution"). The claims-based list is corroborated independently by the presence of the same list of diseases from incident report data. The studies mentioned (Harvard Medical Practice [1984] and Colorado/Utah [1992]) were not included. They fall outside the time window for the systematic review and do not have ED-specific data (so would not have qualified even if they had been in the review's time window).

Commentator & Affiliation	Section	Comment	Response
Anonymous (cont'd)	Methods (cont'd)	(comment above)	It is an interesting question whether root causes (KQ3) identified in malpractice claims might be biased. We have added some text on this issue ("Representativeness of Malpractice Claims Data for Root Causes. It is known that malpractice claims data represent a biased sample of cases, so it is then reasonable to consider whether bias(es) might influence the root causes of diagnostic error identified. As described above, it was clear from ED incident report studies (e.g., Okafor, 2016; Hussain, 2019) that the spectrum of root causes identified is quite similar to that found in ED malpractice claims studies—mostly cognitive errors related to bedside diagnostic decision-making (especially clinical examination, test ordering, or integration of test results into diagnostic reasoning). What is not known is whether both malpractice claims and voluntary incident reports might be biased towards cases with cognitive errors by physicians. This question cannot be easily addressed by retrospective studies relying on chart review, since most potential root causes must be inferred (i.e., they are not actually captured or recorded). Nor can it be addressed by diagnostically oriented, experimental vignette-based studies (which only assess for cognitive errors). To address this question rigorously, one would need a cohort study or clinical trial that prospectively captured all potential root causes and then assessed diagnostic errors and root causes. We found no such studies, so this remains an unanswered scientific question.").

Commentator & Affiliation	Section	Comment	Response
Anonymous	Methods	The review relies on work from certain groups (including what seemingly appears to be the authors themselves) and excluding other groups who have done a lot of work in the area. For instance, while making arguments about focusing on stroke and Big 3 etc. authors consistently refer to their own work heavily to make estimates and suggest implications (many of which are not even directly connected to the review findings). While there is no denying there are several diagnostic errors in the ED in the Big 3 categories, it is also a large chunk of diseases seen in medical practice anyway so I am not sure this classification is as helpful as portrayed. And given so much of cherry-picking, it is not surprising that conclusions are reached about Big 3 'an estimated 69 percent of all ED diagnostic errors resulting in serious misdiagnosis-related harms'. It appears that a lot of systematic bias has occurred while conducting this review especially when other types of issues (such as fractures and other conditions) are inappropriately excluded from emphasis so the authors can quickly focus to make a case for 'vascular, infection, and cancer' as main disease categories.	(1) "work from certain groups" --- No groups of authors were "included" or "excluded" as part of the study methods. All articles for the systematic review and meta-analysis were assessed with respect to the inclusion and exclusion criteria defined *a priori* in the protocol. Authors of studies were not involved in the screening of their own studies for inclusion in the systematic review. In addition, two individuals reviewed each study at the title, abstract, and full text stages and agreement was required to exclude studies from the systematic review. (2) "focusing on stroke" and "refer to their own work" --- The focus on stroke follows from the review's results that identify stroke as the most common cause of serious misdiagnosis-related harms identified in the ED --- thus, it deserves priority from a public health perspective. Furthermore, the number of studies of stroke misdiagnosis (n=18, only one of which involved current study authors) was far greater than any other disease in the review, and none of the studies included in the meta-analysis of stroke error rates in KQ2 was conducted by the authors of the present report. (3) "other ... issues (such as fractures...) are inappropriately excluded" --- Fractures are not part of the Big 3 and were not excluded - they were included in the disease-specific systematic review and discussed in the report. Missed fractures are the most commonly identified diagnostic errors in the ED (in both malpractice claims and incident reports), but they typically cause only low/medium severity harms. In response to this comment, we have done a full review of the fractures in the malpractice claims data source files for KQ1 and recoded fractures (as well as all "other" non-Big 3 diseases) so that they can be appropriately ranked when coded at the same level of granularity. These changes are reflected in Tables 2-4, as well as in the text.

Commentator & Affiliation	Section	Comment	Response
Anonymous	Methods	The authors say they included studies if they were conducted in the United States, Canada, United Kingdom, Western Europe, Australia, or New Zealand. Why only include evidence from 6 developed places?	Based on input from Key Informants and our Technical Expert Panel, who provided input on the scope and protocol, we restricted inclusion to studies conducted in countries with comparable ED care to the US. We have added a sentence to the Methods explaining the rationale ("These nations were chosen in consultation with Key Informants and the TEP to reflect countries with roughly comparable systems of ED care to those found in the United States, in order to maximize representativeness of the final results for US-based ED care. Much less is known about the scope and nature of diagnostic errors in developing nations, but access to basic diagnostic testing resources are very limited in many low- and middle-income countries. As a result, diagnostic delays for life-threatening diseases can be substantial, so studies from these other countries were excluded by design.")

Commentator & Affiliation	Section	Comment	Response
Anonymous	Results	Key point #2 findings rely on 3 studies to make such broad claims but some of these are inappropriate extrapolations. 'A weighted average overall diagnostic error rate of 5.6 percent per ED visit is estimated by combining the error rate among ED discharges (4.1%) from a case-control study at a large university hospital in Spain with the error rate among ED admissions (12.3%) from a rigorous, prospective study at a university hospital in Switzerland.' The Switzerland study is a study on diagnostic discrepancy and not on diagnostic error. It says 'Patients' hospital discharge diagnosis was compared with the diagnosis at hospital admittance through the emergency room and classified as similar or discrepant according to a predefined scheme by two independent expert raters'. The first line of Limitations in this paper literally says 'This study investigated discrepancies in diagnoses, not error, which would require a thorough review of the diagnostic process.' It is unclear why a study that did not even confirm diagnostic error is being used to make such an estimate. In general, such a review should include confirmed diagnostic errors. Cases where there is a discrepant or evolving diagnosis or one where there is an association with a subsequent visit does not necessarily imply diagnostic error. This seriously jeopardizes the estimate.	(1) "inappropriate extrapolations" --- this is discussed in the Applicability section ("Despite sourcing key portions of the data for KQ2 (rates) from a small number of studies conducted in countries outside the United States, we believe the results apply to US-based EDs. Point estimates for overall error and harm rates were drawn from three studies based outside the United States (Canada, Spain, and Switzerland, with a combined n=1,758), but these were the only higher-quality studies found that conducted systematic patient follow-up to minimize under-ascertainment of diagnostic errors. The overall estimated ED diagnostic error rate of 5.7% was far lower than the measured false negative rates for the top serious harm-producing diseases other than myocardial infarction (range 10-56%, Table 9), and of 9 of the 12 disease-specific rates included US-based studies (not pulmonary embolus, meningitis, or pneumonia). The measured overall harm and death rates (though derived from a single, well-designed, prospective Canadian study) triangulate well with data from a nationally representative US-based source (Medicare data on short-term deaths post ED treat-and-release with a "benign" diagnosis). While the referral architecture by which patients attend EDs likely differs across countries (including some included as part of our review), we found no evidence that studies conducted in comparable, disease-specific populations outside the United States had substantively different results than those conducted in US-based EDs. Comparison across studies within each disease did not demonstrate any systematic differences in diagnostic error rates between US-based and non-US-based EDs. The one disease-specific study which included both US-based and European EDs and compared diagnostic performance directly across continents found slightly longer diagnostic delays for aortic dissection patients in North America when compared to Europe; from the list of investigators included in the registry, 12 of 14 North American sites were US-based institutions and the other two were in Canada, while the European sites were from seven countries, including Spain and Switzerland. Thus, there is reason to believe that the error and harm rate estimates are either representative of US ED performance or perhaps even low.");

Commentator & Affiliation	Section	Comment	Response
Anonymous (cont'd)	Results (cont'd)	The study by Calder excluded patients deemed incapable of informed consent (cognitive impairment or major psychiatric illness; critically ill or in distress) or unable to complete 2-week phone follow-up (non-English/French speaker, no telephone, or expected to be unavailable). And it is the only one that provided estimate and others were excluded. The review states 'An estimated overall misdiagnosis-related harm rate of 2.0 percent per ED visit comes from the only prospective study to look at diagnostic adverse events using systematic phone and chart review follow-up of 503 patients both discharged and admitted from the ED. Retrospective trigger-based studies included many more ED visits and often revealed much lower rates, but this was almost certainly due to systematic under-ascertainment.' Why excluded? It appears that the review systematically ignores the retrospective trigger-based studies in favor of one that would give a higher estimate. Why was that done? This is a significant problem because the conclusions of 5.6% and 7 million diagnostic errors are made based on flawed assumptions and weak data.	(2) "The Switzerland study is a study on diagnostic discrepancy and not on diagnostic error." --- we used the NAM definition of diagnostic error (which does not require a process failure --- as noted in the Background, "The National Academy of Medicine (NAM) defines diagnostic error as "the failure to (a) establish an accurate and timely explanation of the patient's health problem(s) or (b) communicate that explanation to the patient."). Notably, this definition (which is used in this report) does not require a care process failure (e.g., a specific clinical reasoning "mistake" on the part of an individual clinician) and is agnostic with respect to any resulting harms or their preventability."), so the Swiss study was, in fact, a study of diagnostic error by this definition, regardless of the terminology used in the paper by the original study authors; (3) Calder study exclusions - correct, these exclusions are listed in the footnotes to Table 8, so they are available to readers to judge the generalizability of results. However, it is not likely that diagnostic error rates would be substantially lower among patients who are cognitively impaired or unable to communicate because of a language barrier (which is noted in Okafor, 2016 as one of the causal risk factors for misdiagnosis);

Commentator & Affiliation	Section	Comment	Response
Anonymous (cont'd)	Results (cont'd)	This bias gets worse for making mortality estimates on page 27 when they use one death for an estimate: 'An estimated overall misdiagnosis-related death rate of 0.2 percent per ED visit comes from the only prospective study to look at diagnostic adverse events using systematic phone and chart review follow-up of 503 patients both discharged and admitted from the ED. This estimate is based on just a single death' Again the retrospective studies are ignored seemingly because the rate was much lower. 'Misdiagnosis-related deaths per ED visit were reported in three of four retrospective studies,24, 69, 136 ranging from 0 to 0.007 percent…' In sum, the review reflects very obvious bias with which review ignores certain studies in favor of others to support a certain claim.	(4) retrospective studies were excluded --- Studies were not excluded from the review based on particular study designs or results. As noted in the protocol, "We conducted meta-analyses when there were sufficient data (i.e., at least two studies) and studies were sufficiently homogenous with respect to key variables (e.g., population characteristics, condition, provider type, and data source/study design)." It is inappropriate to pool studies of different designs (e.g., those with different sampling frames). The retrospective studies available to address the disease-agnostic error and harm rates used very different inclusion criteria (e.g., only included patients who returned to the ED, rather than consecutive patients seen in the ED). The details of these design differences are discussed in the sub-section of KQ2 entitled "Per-Visit Overall ED Misdiagnosis-Related Harm Rates" (see 7 consecutive paragraphs beginning with "There were four retrospective studies that reported overall per-visit harm rates..."). The argument is summarized again in the Frequently Asked Questions (FAQ) document (Q#51, Newman-Toker DE, Peterson SM, Badihian S, et al. Frequently Asked Questions for "Diagnostic Errors in the Emergency Department: A Systematic Review". Open Science Framework, 2023. DOI: https://doi.org/10.17605/OSF.IO/B7XVM). (5) "This bias gets worse for making mortality estimates" --- Checks on the validity of these estimates, including mortality, are described in section KQ2a (see especially section entitled "Plausibility of Mortality Estimates from Higher Quality Studies") and KQ2c (see especially paragraph beginning "Although these estimates may seem high, they are on par with what has been estimated for harms from inpatient diagnostic error (250,000 harms out of 36 million hospitalizations), based on systematic review data...").

Commentator & Affiliation	Section	Comment	Response
Anonymous	Results	On page 15, authors discuss about most common symptoms associated with misdiagnosis 'Data were sparse with respect to the most commonly misdiagnosed clinical presentations (symptoms, signs, or syndromes).' Unclear how authors are able to comment on these symptoms when they handpicked certain diseases to look at and used malpractice claims data for most of the work. Moreover, the denominator for such a prevalence estimate should be the total number of presentations to the Emergency Department in which a subsequent diagnosis is a possibility. When you look at 'dizziness', a missed stroke is pretty low about 0.2% (see Atzema paper https://onlinelibrary.wiley.com/doi/10.1002/ana.24521) This type of methodology over-estimates diagnostic error rates. '45,000 to 75,000 missed strokes', refers to a perspective type paper.	(1) "unclear how authors are able to comment on these symptoms" --- It was prespecified in our protocol that we would assess the symptoms, signs, and conditions most frequently associated with diagnostic error in the ED. Diseases were selected with input from Key Informants and the Technical Expert Panel, and symptoms were abstracted in relation to these diseases.; (2) "used malpractice claims for most of the work" --- almost none of the studies cited with respect to symptoms were malpractice studies (for example, see KQ3, Illness Characteristics, in which 120 studies are cited); (3) "the denominator... should be the total number of presentations to the ED" --- when studies reported on a symptom-based cohort (e.g., Dubosh, 2020), we reported the denominator in this way (e.g., false omission rate); when studies used a disease-based framework, we reported studies in that way (e.g., false negative rate) – doing anything else would have been both methodologically inappropriate and, almost invariably, not mathematically possible [given the nature of the data that were reported in various studies];

Commentator & Affiliation	Section	Comment	Response
Anonymous (cont'd)	Results (cont'd)	(comment above)	(4) "This type of methodology over-estimates diagnostic error rates" --- it is to be expected that (a) the proportion of missed strokes that presented with dizziness (denominator is all strokes) or (b) the proportion of dizzy-strokes that are missed (denominator is a subset of strokes) might be substantially different than the proportion of patients with dizziness who suffer harms from missed stroke (denominator is all with dizziness symptoms) --- this fact is unsurprising, but it is a large part of the reason why there are apparent discrepancies in "diagnostic error rates" as reported in the literature, and we are careful throughout the report to clarify when different denominators are being used, and not to inappropriately combine data that use different denominators (see Methods paragraph beginning "The error and harm rates, which are the focus on KQ2, may have been expressed differently in different studies...") and to discuss how different denominators impact inferences (see the "Implications for Operational Quality Measurement and Benchmarking" section, sub-heading "Differences in causal inferences based on different denominators"); (5) "'45,000 to 75,000 missed strokes', refers to a perspective type paper" --- true, that was from an editorial accompanying the Atzema paper cited by this commenter, but that editorial, in turn, cites multiple primary data studies and explains in detail the evidence-based rationale for the numerical estimate.

Commentator & Affiliation	Section	Comment	Response
Anonymous	Discussion	Authors propose SPADE methodology but do not note its application or limitations. It would be helpful to note which and how many institutions are currently using this methodology in routine clinical care for operational measurement and benchmarking as proof of concept and what types of actions are being taken on patients who are being identified through this method. Further, this method when validated for stroke via rigorous medical record reviews produced a PPV of only 33% so how this can be used for benchmarking should be clarified https://academic.oup.com/jamia/article-abstract/28/10/2202/6324038	(1) "do not note its application or limitations" --- We address application of SPADE in the section entitled "4. Approaches to measurement at the institutional level." stating, "No single measurement method or individual measure will suffice. A "portfolio" approach is needed. A one-size-fits-all approach is unlikely to be equally appropriate for all institutions. Offered below are a few different ways that an institution might choose to approach measuring diagnostic errors." We address limitations in the Discussion section when discussing use of SPADE for measurement ("However, SPADE relies on detecting adverse events. From the studies we identified, these are relatively infrequent (typically less than 1 percent of treat-and-release cases), so stable measurement generally requires thousands of encounters. That means that at a medium to large-sized ED, relatively common symptoms (e.g., abdominal pain, chest pain, dizziness, headache, back pain) can be mined using SPADE for misdiagnosis-related harms linked to more common dangerous diseases such as stroke, myocardial infarction, sepsis, or pneumonia using a rolling 6- to 12-month window. Smaller hospitals or rarer diseases generally require longer assessment time windows." and "Missed cancer (including lung cancer) may require alternative monitoring methods, since the temporal risk profile of adverse events after a lung cancer misdiagnosis are very different than those after a missed vascular event or infection, making it less readily amenable to current SPADE methods."; (2) "currently using this methodology" --- we agree it would be helpful to know how many institutions were using routine operational measures of diagnostic error, in general (SPADE or otherwise), but, to our knowledge, no systematic data on this topic are available; (3) "how this can be used for benchmarking should be clarified" --- We devote an entire sub-section of the Discussion to how SPADE can be used for benchmarking "5. High-stakes measurement for accountability, payments, and national benchmarking."

Commentator & Affiliation	Section	Comment	Response
Anonymous (cont'd)	Discussion (cont'd)	(comment above)	The public commenter's statement that chart review-based methods are "rigorous" and, by implication, therefore a good gold/reference standard by which to gauge the accuracy of SPADE is inconsistent with available evidence about the quality and consistency of chart reviews for assessing medical error (see PMID: 11466119 and Background section of PMID: 29358313). First and foremost, the Vaghani et al. study cited by the commenter did not validate "a PPV of only 33%" --- in fact, Vaghani et al. estimated there was no missed opportunity in just 23% of reviews, while the other 77% of cases were deemed missed opportunities, possible missed opportunities, or inconclusive, with the latter two categories reflecting data missingness in charts or other reasons why the chart review was inadequate. Second, SPADE does not speak to whether there was a process failure, only to whether there was an adverse event, and adverse events were not defined in the Vaghani study, so it cannot reasonably be used to judged SPADE's accuracy. Third, SPADE is a direct measure of adverse events that, unlike chart review methods, does not require subjective judgments that can be influenced by data missingness, hindsight bias, and differential domain knowledge among human raters. In multiple studies by multiple groups, SPADE and related methods have clearly shown statistically valid variation across hospitals, hospital types, and provider types (e.g., PMID: 32701479, 17112926, 17322078, 28344918, 29540019, 34147048) and can be completed using H-CUP or Medicare data, making it a sensible choice for state-level or national-level benchmarking across US hospitals at minimal cost (compared to routine chart review). To make this last point more clearly, we have added the following text to the benchmarking section ("..., which have been shown to vary substantially by hospital (e.g., for acute myocardial infarction, where misdiagnosis-related adverse event rates varied 3.3-fold from 0.6% to 1.9% across individual EDs, $p<0.001$) and permit observed minus expected analysis to detect statistically valid excess adverse events above the base rate.").

Commentator & Affiliation	Section	Comment	Response
Anonymous	Discussion	For key question 3 the review states "The most robust data about overall root causes came from a large, US-based malpractice claims study and a large, UK-based incident report study." It is surprising that such a systematic review could not find other ED studies that have evaluated cognitive error. This implies there are no other methods or studies that have provided data on cognitive error in the Emergency Department which sounds so hard to believe just at face value. Even if only malpractice claims were included the Kachalia study they cite would be equally or more robust. I would recommend against making assertions that cannot be supported by the literature.	There were many studies that addressed cognitive errors, but very few with identical categorization schemas that permitted aggregation. Kachalia et al. studied cases from 1979 and 2001 and did not break out cases from 2000-2001, so the study was excluded from the analysis, per the pre-specified methods as described in our protocol. Nevertheless, Kachalia gives us a good reference point for the robustness of the two large, cited studies in this section. The two largest studies described had 2,273 ED diagnostic error claims and 1,577 incident reports with causes identified. By contrast, Kachalia et al. (one of the larger malpractice studies other than the one shown in Figure 16), reported on just 122 cases (5% of the sample shown in Figure 16 and more than 30-fold fewer cases than the malpractice and incident report studies combined). Furthermore, Kachalia et al. found essentially the same thing (failure to order tests in 58%, failure to perform an adequate history and physical exam in 42%, failure to correctly interpret a test in 37%, and failure to order an appropriate consultation or make a referral in 33% --- and far lower rates of non-cognitive errors).

Commentator & Affiliation	Section	Comment	Response
Anonymous	Discussion	Discussion is too long and many points in discussion and policy changes are mostly unrelated to the purpose of the review and are tangential assertions by the authors. Some of these are best classified as personal opinions not implications and findings that are supported by this review. The key finding is we do not have an accurate estimate and better studies are needed. The policy suggestions of dashboards when most cannot even identify diagnostic errors is premature. Several untested methods are being proposed that sounds very overpromising for an AHRQ review and seems beyond scope. Perhaps this large recent review can provide a reality check- Dave BMJQS https://qualitysafety.bmj.com/content/31/4/297	(1) As described, suggested policy changes are ones that should be considered, rather than strong policy recommendations ("Policy changes to consider based on findings from this review include..."). Furthermore, we believe that these suggestions derive directly from the report's findings (rather than being "unrelated" or "tangential"): "(1) standardizing measurement and research results reporting to maximize comparability of measures of diagnostic error and misdiagnosis-related harms" - this derives directly from the lack of standardized measurement of diagnostic error and harms identified by the systematic review; "(2) creating a National Diagnostic Performance Dashboard to track performance (analogous to the Dartmouth Atlas Project for utilization of healthcare services)" - this derives from the lack of adequate national benchmarking and lack of comparability of measurement across EDs identified in this systematic review; and "(3) using multiple policy levers (e.g., research funding, public accountability, payment reforms) to push for the rapid development and deployment of solutions that address this major patient safety and quality problem" - this derives directly from the overall public health scale/scope of the problem identified by the review. (2) The systematic review of interventions to address diagnostic error (mentioned by the commenter) did not include policy changes among their inclusion criteria, so is not germane to this issue.

Commentator & Affiliation	Section	Comment	Response
Anonymous	Conclusion	Reiterating that in my opinion, this systematic review does not do what it was supposed to do. Findings are not valid and are highly biased. Handpicking areas and diseases to focus on prior to the review, using largely "numerator-only" such as malpractice claims to make assessments and answering key questions, cherry-picking selected studies while ignoring others to make estimates, and relying on studies that did not even include diagnostic error are concerns that are scientifically significant. All of these are "fatal" methodologic flaws and introduce bias that cannot be easily fixed. In addition, the implications and policy section is largely disconnected from the review and seems to be focused on furthering a highly selective and narrow personalized agenda.	The methods for this systematic review were described *a priori* in our protocol. The methods were developed with input from 14 individuals included in either our Technical Expert Panel or as Key Informants; members included those with expertise in emergency medicine, emergency nursing, patient safety and quality, epidemiology and a patient with lived experience with diagnostic error in the ED. The review questions were also posted for comment and groups such as the American College of Emergency Physicians offered their input—they made helpful methodological suggestions that were incorporated into the design, including assessing the literature for the impact of both fixed and dynamic systems factors in diagnostic error (October 23, 2020). The findings from the systematic review are based on the best available evidence, and we hope that the review will help to identify opportunities to improve the quality and quantity of high-quality evidence about this important issue.
Anonymous	References	References selectively include work from certain groups (including what seemingly appears to be the authors themselves) and exclude other groups who have done a lot of work in the area	The methods for this systematic review were described *a priori* in our protocol, with many processes in place to safeguard against bias. Inclusion and exclusion criteria were determined *a priori* and study authorship was not part of the eligibility criteria for studies. Consistent with the mitigation plan to avoid bias, study authors were not involved in determining eligibility of their own studies for inclusion in the systematic review.
Ritu Agrawal	Methods	Hypovolemic shock occuring in pregnancy is missing	Hypovolemic shock in pregnancy was not identified in the prespecified list based on preliminary literature searching and input from Key Informants and the Technical Expert Panel. Other conditions related to pregnancy (ectopic pregnancy and pre-/eclampsia) were included.

Commentator & Affiliation	Section	Comment	Response
Ritu Agrawal	General	Pregnancy related conditions are missing out of the report. I feel this should be part of it e.g. abortion related mis-diagnosis, ectopic or Post Partum haemorrhage etc.	Ectopic pregnancy and pre-/eclampsia were explicitly included in the search and are reported on. Ectopic pregnancy appears in Table 2. Unfortunately, there were no rate-based studies of these conditions to be included in the review, as stated at the end of the KQ2b section ("We did not find any studies meeting our inclusion criteria that reported on the ED diagnostic error rate for endocarditis, necrotizing enterocolitis, sudden cardiac death, arrythmias, congenital heart disease, ectopic pregnancy, or pre-eclampsia/eclampsia.") We have also added language to the limitations section ("However, because of the constrained focus on the most common conditions, we do not have data on misdiagnosis of less common conditions that may nevertheless be of importance to ED clinicians (non-accidental trauma, necrotizing fasciitis, compartment syndrome, brain tumors, obstructive hydrocephalus, ovarian torsion, post-partum hemorrhage, etc.); this is a limitation.").
Linda Estep	Results	Were any evaluations done on error rates for people of color or other marginalized groups (LGBTQIA, homeless mentally ill, etc.) segments of those in the study?	Yes, the impact of race and ethnicity was examined and is described under "Patient Characteristics" in KQ3. Other factors are identified as gaps for future study, and we have expanded the text in this section to explicitly list the groups mentioned by the commenter, among others ("Other patient characteristics reflecting marginalized status (e.g., members of religious minorities; lesbian, gay, bisexual, transgender, and queer [LGBTQ+] persons; persons with disabilities; persons who live in rural areas; and persons otherwise adversely affected by persistent poverty [including homelessness] or inequality) or the presence of marginalizing co-morbidities (e.g., mental health or substance use disorders or obesity) that may increase the risk of diagnostic error are understudied and deserve further equity-related research.")

Commentator & Affiliation	Section	Comment	Response
Ryan Radecki	General	My comments on this draft report do not fit tidily into the above structured format, as my general concern is this report falls quite short in its goal of providing a reliable evaluation of diagnostic error in the Emergency Department. The foundational data relied upon by these draft authors is of such low reliability the conclusions remain grossly unsupported. Even though this draft report does not have authors listed, this report relies unusually upon repeated citation of a certain author's work. Due to this apparent professional bias, certain diagnoses and policy suggestions are provided undue attention. This report requires wholesale revision.	The methods for this systematic review were described *a priori* in our protocol, with many processes in place to safeguard against bias. Inclusion and exclusion criteria were determined *a priori* and study authorship was not an inclusion or exclusion criterion. As a part of our processes to safeguard against bias individuals authoring studies were not involved in screening of their own studies for inclusion and exclusion. Finally, all authors of the report recused themselves from risk of bias assessment and data extraction of their own studies. We have added language to the front matter describing the processes to mitigate potential conflicts of interest by the authors of the report. We also do not agree with the assertion that the report "relies unusually" upon a "certain author's work". For instance, only 7 studies (2.5% of 279 included studies in the review) had Dr. Newman-Toker as an author.

Commentator & Affiliation	Section	Comment	Response
Ryan Radecki	Methods	The discussion around Key Question 1 cites Newman-Toker work in which the "Big Three" classes of diagnoses are emphasized as those with the greatest magnitude for harm. In fact, this citation cannot reach any other conclusion, as the assumption underlying the methods for the cited work is these three classes contained the greatest likelihood for harm, and no other diagnostic classes were examined. The authors of this draft cite multiple other collections of diagnostic error from overseas, but these collections of errors â€“ fractures, abdominal pain, pregnancy complications, etc. â€“ are inappropriately discarded to focus on vascular, infection, and cancer. Worse still, the authors rely heavily on this data set for frequency estimates despite being derived from closed malpractice claims. There is no reliable basis for claiming the frequency of tort filed as a surrogate for the rate of diagnostic error. This bias makes the choice of this data set inappropriate, particularly as it is obviously discordant from the other data reviewed.	(1) "this citation cannot reach any other conclusion... no other diagnostic classes [other than the Big Three] were examined" --- In the Newman-Toker malpractice claims study, all classes of disease were included and assessed (not merely the “Big Three”); however, the rankings were potentially influenced by the Big Three structure (i.e., subcategories of "Other" were not broken down, so, in the original paper, they were not ranked relative to subcategories of vascular events or infections, for instance). In response to this comment, and to determine whether this was an issue in the present analysis, we re-analyzed the HCUP-CCS categories so that the rankings could be re-assessed. We completed a full review of the fractures in the malpractice claims data source files and recoded fractures as well as all "other" (non-Big 3) diseases so that they could be appropriately ranked when coded at the same level of granularity. The resulting changes are now reflected in Tables 2-4, as well as in the text. In response to this comment, we also added a breakdown of conditions by organ system (Table 4) in addition to by "Big Three" categories. (2) "There is no reliable basis for claiming the frequency of tort filed as a surrogate for the rate of diagnostic error" --- We agree, and the report did not do this---none of the incidence/rate questions addressed in KQ2 were answered using malpractice data. Issues of representativeness of malpractice claims data for KQ1 (frequency distribution of diseases causing serious harms, NOT prevalence) are addressed in "Representativeness of Malpractice Claims Data for Disease Distribution" and for KQ3 are addressed in "Representativeness of Malpractice Claims Data for Root Causes". To address any confusion over data sources with respect to specific KQs, we have added an Appendix on Data Types/Sources (see Appendix Table A-1).

Commentator & Affiliation	Section	Comment	Response
Ryan Radecki	Methods	The framing of the frequency of missed diagnosis is also fundamentally incorrect. From a healthcare quality perspective, the denominator is the total number of presentations to the Emergency Department in which a subsequent diagnosis is a possibility. Relevant to this draft, take "dizziness" for example. Repeated work has shown the frequency with which a patient presenting to the Emergency Department with dizziness is erroneously given a non-stroke diagnosis (e.g., a missed stroke) is approximately 0.2% (ex https://doi.org/10.1002/ana.24521). However, these authors frame the missed diagnoses by looking backwards from cohorts of known final diagnoses, presenting in terms of the frequency with which certain presenting symptoms were associated with a delayed or missed diagnosis. Focusing on the "false negative rate" rather than the "false omission rate", the authors of this draft exaggerate the magnitude of the problem of misdiagnosis.	Certainly it is true that false negative rates and false omission rates will differ in the vast majority of cases. However, one is not "correct" and another "incorrect" --- they are equally valid descriptors of different denominator populations, and, depending on study design, may reflect either the diagnostic error rate or the misdiagnosis-related harm rate. Importantly, we did not "focus" on one over the other. We reported on the data provided in the studies. There were many more studies reporting false negative rates than reporting on false omission rates (see Table 9, which reports both false negative rates and false omission rates, including the false omission rate measured for dizziness as 0.2%, as noted by the commenter). As a general matter, for undifferentiated symptoms that are infrequently due to dangerous diseases, the false omission rate will be lower, especially when the method of ascertainment relies on return visits... which reflects only the subset of diagnostic errors with adverse events, rather than all diagnostic errors (since some patients with incorrect diagnoses do not return and, therefore, go unascertained). Since all of the studies examining false omission rates relied on hospitalizations for a dangerous disease after a benign discharge, these are reported on in relation to the misdiagnosis-related harm rate, not the diagnostic error rate.

Commentator & Affiliation	Section	Comment	Response
Ryan Radecki	Results/methods	The results generated by the draft authors analyses in response to Key Question 2 are, in no uncertain terms, farcical. Their determination of the rate of diagnostic error depends, effectively, on an arbitrary blend of rates from a grand total of two studies. The first of these studies is the Nunez study, looking at two cohorts of 250 patients discharged from the ED in Spain. The draft authors cite a miss rate of 4% (10/250) for the "non-returns" in this study, but it appears the actual number in the paper for "diagnostic error" in "non-returns" is 1.2% (4/250). As "non-returns" make up the bulk of their study cohort (~32,000 patients), this effectively defines the rate of diagnostic error. Furthermore, the gold standard for the correct diagnosis, and thus source for the determination of diagnostic error, is a subsequent diagnosis provided by a primary health center. This definition of diagnostic error, even as a surrogate, cannot be considered a validated and accurate method for such.	(1) Nunez study --- In Nunez (Table 2), the diagnostic error rate was 20% of 250 unscheduled returns and 4% of 250 who did not return but were followed up. It is understandable that the commenter might be confused about the numbers from the Nunez study, since it requires careful attention to the footnote to Table 2 to identify that the numbers presented are percentages not n's. The final estimate of 4.1% from the report is a blended rate (weighted average) of the two values, as described in the text ("Thus, diagnostic errors were 5-fold enriched among patients with 72-hour returns, but because the unscheduled return rate was just 0.8 percent, the estimated total diagnostic error rate for the entire ED population was 4.1 percent. This likely represents a "floor" (minimum) rate estimate because diagnostic errors were based solely on chart review and not systematic patient follow-up."). (2) Hautz study --- The Swiss study was a strong, prospective study with systematic ascertainment of the diagnostic errors, that reported higher mortality among patients who were misdiagnosed ("They found diagnostic differences in 42 percent of cases (n=319 of 755) and considered meaningful discrepancies in 12 percent of cases (n=93 of 755). Although the authors demurred labelling these as errors (focusing on "error" as a process failure), these events meet the National Academy of Medicine definition of a diagnostic error, regardless of whether an explicit, preventable failure occurred during the diagnostic process. Diagnostic errors were associated with longer hospital stay (mean 10.3 versus 6.9 days; Cohen's d 0.47; 95 percent confidence interval 0.26 to 0.70; P = 0.002) and increased patient mortality (8.6 percent [n=8] versus 3.8 percent [n=25]); odds ratio 2.40; 95 confidence interval 1.05 to 5.5 P = 0.038). Note that no post-hospital follow-up was performed, so the authors concluded that their estimates were likely minimum estimates (i.e., some additional diagnostic errors were presumably not captured by the inpatient team and therefore unaccounted for in the study results).")

Commentator & Affiliation	Section	Comment	Response
Ryan Radecki (cont'd)	Results/methods (cont'd)	The second of these studies involves patients admitted to an internal medicine service in Switzerland. This, by definition, excludes large swathes of emergency department patients admitted to other services, limiting the accuracy and generalizability of its measurement. This study again relies upon coded diagnoses as a surrogate for diagnostic error. Given the general milieu of an internal medicine service, it is likely many of these patients classified as "diagnostic error" are better classified as "diagnostic uncertainty", in which an inpatient evaluation is required as the scope of the emergency department timeframe is exceeded. A lack of structured chart review severely limits the accuracy of their point estimate. The draft authors then merge these results into their own estimate, 5.6%, and use this to define their topline results and conclusion of 7,300,000 instances of diagnostic error in the Emergency Department in the United States. The underlying data is so profoundly weak it is grossly unacceptable to generate and promulgate such a number as a definitive finding.	(3) Topline results --- The data were the best available and in the final report we have provided additional text around the limitations of the primary studies. In addition, the final report acknowledges these limitations and calls for additional rigorous studies so that we can have the most accurate estimate of diagnostic error in the ED ("Future research should emphasize areas in which data are suboptimal or lacking. For decision making in the United States, overall diagnostic error and harm rates should be confirmed in U.S.-based studies using rigorous, prospective methods.") We describe in the text the rationale for why we believe these numbers are valid estimates of the total diagnostic error rate ("... we estimate a weighted average overall diagnostic error rate of 5.7 percent per ED visit by combining the error rate among ED discharges (4.1%) from a case-control study at a large university hospital in Spain with the error rate among ED admissions (12.3%) from a rigorous, prospective study at a university hospital in Switzerland. The representativeness of this estimate is uncertain, but the figure is not outside the range expected based on disease-specific error rates found in KQ2b, which range from 1 (fractures, myocardial infarction) to 56 percent (spinal abscess). Additionally, the 4.1 percent estimate for the ED diagnostic error rate is correctly positioned within the spectrum of error/harm frequencies—diagnostic errors among admitted "non-specific" symptom cases (54%) > diagnostic errors among admitted patients (12%) > diagnostic errors among treat-and-release discharges (4%) > diagnostic errors resulting in adverse events (2%) > diagnostic errors resulting in serious harms, including death or permanent disability (0.3%). Finally, the overall error rate of 5.7% is comparable to that found in rigorous US-based studies of other frontline care settings (e.g., 6.3% overall diagnostic error rate in US-based primary care clinics). Thus, we believe it is appropriate to report this result."). To address issues of uncertainty in the data and small numbers of (rigorous) studies on which we rely, we have added additional text to the Executive Summary and the Conclusions ("Our review findings are tempered by limitations in the underlying evidence base, including issues linked to data sources, measurement methods, and causal relationships.") and we have added plausible ranges around the point estimates for extrapolated numbers.

Commentator & Affiliation	Section	Comment	Response
Ryan Radecki	Results/methods	The draft authors perform a similarly inappropriate generalization with respect to misdiagnosis-related harm rates. For patients discharged from the ED, trigger-based retrospective studies are discarded in favor of a single prospective study by Calder. This study enrolled patients registered to the high-acuity portion of an ED in Canada, in which a misdiagnosis-related adverse event rate of 2.0% is witnessed, as well as a 0.2% rate of death. Despite being, as these authors state, "217-fold higher than the weighted mean from the three retrospective studies", it is incredulously "not far from what appears to be a plausible range". The authors make several other inappropriate leaps of extrapolation from other data sets, including the Swiss inpatient study noted above, to support this claim. Despite these limitations, the draft authors unflinchingly use the 2.0% and 0.2% rates as their topline results, escalating their flawed analysis by applying these numbers to their estimated error rate of 7,300,000, finding 2,600,000 diagnostic adverse events and 260,000 misdiagnosis-related deaths. One person dies of an aortic dissection in Ottawa, and thus, absurdly, there are 260,000 annual deaths in the United States.	To address issues of uncertainty in the data and small numbers of (rigorous) studies on which we rely, we have added additional text to the Executive Summary and the Conclusions ("Our review findings are tempered by limitations in the underlying evidence base, including issues linked to data sources, measurement methods, and causal relationships."). We also added plausible ranges around the point estimates for extrapolated numbers. And we have since prepared an Addendum to the original report that addresses statistical concerns by using Monte Carlo simulations to combine results across studies while generating true, statistically valid 95% confidence intervals rather than plausible ranges. As explained in the text, retrospective studies were not excluded from the report but were excluded from specific analyses because they used very different sampling frames. The retrospective studies available to address the disease-agnostic error and harm rates used very different inclusion criteria (e.g., only included patients who returned to the ED, rather than consecutive patients seen in the ED). The details of these design differences are discussed at some length in the sub-section of KQ2 entitled "Per-Visit Overall ED Misdiagnosis-Related Harm Rates" (see 7 consecutive paragraphs beginning with "There were four retrospective studies that reported overall per-visit harm rates..."). The argument is summarized again in the Frequently Asked Questions (FAQ) document (Q#51, Newman-Toker DE, Peterson SM, Badihian S, et al. Frequently Asked Questions for "Diagnostic Errors in the Emergency Department: A Systematic Review". Open Science Framework, 2023. DOI: https://doi.org/10.17605/OSF.IO/B7XVM).

Commentator & Affiliation	Section	Comment	Response
Ryan Radecki (cont'd)	Results/methods (cont'd)	(comment above)	The Calder 2010 study used rigorous methods that, nevertheless, undercounted misdiagnosis-related harms --- they counted management errors as a consequence of diagnostic inaccuracy as treatment errors --- whereas in our definition of misdiagnosis-related harms used in this report, these should properly have been counted as diagnostic error-related; accordingly, the estimates from this study are, if anything, low. While the overall serious harm rate (morbidity/mortality) in the original report was nominally based off the result from the well-designed Calder study (in which just one patient died), the measured rate of 0.2% was in keeping with the available evidence synthesized in the report, including deaths from the two other prospective, disease-agnostic studies (n=36) (~0.3%). In KQ2a, we devote several paragraphs to triangulating evidence from multiple sources to bolster the fact that this value is likely close to the true value. These were assembled in response to this (and similar critiques) as a section entitled "Plausibility of Mortality Estimates from Higher Quality Studies." The Calder study was not used in isolation but could not easily be mathematically combined with the other studies (showing the similar ~0.3% mortality rate) because of the way in which the data had been presented by the original studies. As noted above, we have since prepared an Addendum to the original report that addresses these statistical concerns by using Monte Carlo simulations to combine results across studies while generating true, statistically valid 95% confidence intervals rather than plausible ranges. The result of this new analysis that combines studies meta-analytically was slightly higher estimates of mortality (0.22% instead of 0.2%).

Commentator & Affiliation	Section	Comment	Response
Ryan Radecki	Methods	The draft authors approach to Key Question 3 is also foundationally flawed. Their analysis of the frequency of causes of error relies, again, on an analysis of tort claims and the data included therein. There is no validation for use of tort claims as a surrogate for the generalizable frequency of sources of diagnostic error.	A portion of KQ3 does use malpractice claims data and incident report data to assess causes (KQ3a, focused on "root causes" which are often only reported in malpractice or incident report studies). However, the largest section of KQ3 does NOT use malpractice claims (KQ3d, focused on risk factors for diagnostic error). Even for KQ3a, there is nothing surprising about the list of causes derived from malpractice claims, which, as noted in the report, is quite similar to what was found in the largest ED-based incident report study (Hussain, 2019). Furthermore, although detailed categorization schemas vary widely by study author, the same basic distribution of error causes (dominated by cognitive errors in bedside decision-making) is found in most studies of diagnostic error across clinical settings, including outside the ED (e.g., PMID: 19901140, 16009864, 23440149). Nevertheless, it is possible that some forms of reporting bias are potentially at play. Therefore, a section entitled "Representativeness of Malpractice Claims Data for Root Causes" has been added in response to this comment. Lastly, the data were the best available, and in the final report we have provided additional text around the limitations of the primary studies ("Our review findings are tempered by limitations in the underlying evidence base, including issues linked to data sources, measurement methods, and causal relationships.").

Commentator & Affiliation	Section	Comment	Response
Ryan Radecki	Limitations Section	The limitations section fails to acknowledge the gross inadequacy of the evidence base. After performing a systematic review in which only two supposed high-quality studies were found with which to estimated the frequency of diagnostic error, this synthesis is clearly not a "gap filled". The conclusion "missed vascular events and infections" are the principle harms comes directly from a citation in which all other harms were, by definition and initial assumption, excluded. Very clearly, the takeaway with respect to Key Question 1 is a mammoth hole in the evidence base regarding the scope of diagnostic error in the Emergency Department. Similarly, the "gaps filled" for Key Question 2 are based on a synthesis of data framing a "necessarily imprecise" estimate from a single study. There is no face validity to the draft authors' "meta-analytic-supported conclusion of increased mortality". Again, we see an absolute absence of useful data to inform future direction and policy.	(1) "gross inadequacy" --- We sought the best available evidence and acknowledged the limitations. We have added additional text to the Executive Summary and the Conclusions ("Our review findings are tempered by limitations in the underlying evidence base, including issues linked to data sources, measurement methods, and causal relationships."). We called for more research. (2) "missed vascular events and infections..." --- It is not correct that this comes from a single citation --- it is synthesized from two very large studies (one US malpractice claims based and the other UK incident report-based) that represented 78% of the meta-analytic cases. Also, the same essential findings were present in the other (smaller) studies identified ("There were nine studies that addressed KQ1a directly for all diagnostic errors, reporting on a total of 5,817 diagnostic errors. Four studies were malpractice claims-based and five were incident report-based. The two largest studies, one a large, United States-based review of a national malpractice claims database (Newman-Toker, 2019) and the other a large, United Kingdom-based review of a national incident reporting system (Hussain, 2019) together represented 78 percent of diagnostic error cases (n=4,561 of 5,817). These two studies organized their categories in similar enough fashion to present results together (Table 2)."); nor is it true that from the malpractice claims study "all other harms were, by definition and initial assumption, excluded" (see detailed response to a previous issue raised by this commenter beginning with "The discussion around Key Question 1 cites Newman-Toker work in which the "Big Three" classes of diagnoses are emphasized..."). (3) "the takeaway ... is a mammoth hole in the evidence base regarding the scope of diagnostic error in the ED" --- We disagree with the "mammoth" characterization, but we agree that more research is needed to refine these estimates using rigorous methods, as we noted in the Conclusions ("Importantly, large, prospective studies are needed to validate current diagnostic error rate estimates, as well as to help develop valid proxy measures that are more readily and routinely acquired for operational measurement.").

Commentator & Affiliation	Section	Comment	Response
Ryan Radecki (cont'd)	Limitations Section (cont'd)	Likewise, the conclusions made with respect to the cognitive errors and root causes appears based on review solely of tort claims, rather than high-quality data examining emergency department operations and cognitive biases. The field of research into diagnostic error contains many methods with potentially rich contribution to further insight into the cause of cognitive error in the Emergency Department, many of which are touched upon by these draft authors. The takeaway with regard to Key Question 3 ought to be to further pursue better understanding of those errors, including which can be realistically mitigated. Further, the idea errors can be reduced by simply addressing "problems in fundamental bedside diagnostic skills and clinical reasoning" is facile. The scope of knowledge required to practice safely in Emergency Medicine is vast, and necessarily emphasis on a single diagnosis â€“ or atypical presentations of such diagnosis â€“ simply changes the direction of the cognitive biases present. These simplistic proposals to improve diagnostic expertise require themselves an entire line of prospective research into methods for diagnostic- and decision-support to determine effectiveness and unintended harms.	(4) "The takeaway ... ought to be to further pursue better understanding of those errors, including which can be realistically mitigated." --- We agree with the commenter and have added text to this effect to the Conclusions ("A key focus of research should be to define symptoms and diseases for which diagnostic errors and associated harms can realistically be mitigated and to measure the real-world impact of interventions and strategies in reducing these errors and harms."); (5) ""proposals to improve diagnostic expertise require themselves an entire line of prospective research into methods for diagnostic- and decision-support to determine effectiveness and unintended harms" --- We agree with the commenter and have added text to this effect to the Conclusions ("All of these solutions should be subjected to rigorous outcomes research to assess any benefits to improved diagnosis or unintended consequences (e.g., test overuse)..").

Commentator & Affiliation	Section	Comment	Response
Ryan Radecki	Conclusion	The concluding policy statements offer many potential suggestions, and broadly collate many different methods, for monitoring diagnostic errors, types of triggers for review, and the necessary compromises between laborious manual/NLP structured review and imprecise retrospective reporting. It is clear from the poor evidence supporting the answers to Key Question 1 and 2, the most important next step, from a national patient safety perspective with regard to diagnostic error in the Emergency Department, is to fill the gaps with large-scale studies improving upon the methods of the handful of studies relied upon by the draft authors.	We agree that more systematic measurement of diagnostic errors and harms is a critical next step, which is why the first of our three main policy considerations from the report directly address this issue ("(1) standardizing measurement and research results reporting to maximize comparability of measures of diagnostic error and misdiagnosis-related harms") and the second flows immediately from the standardization of measurements ("(2) creating a National Diagnostic Performance Dashboard to track performance").
Ryan Radecki	methods/references	Lastly, circling back to the concern for professional bias. The nature of the reliance upon the Newman-Toker work regarding tort claims restricted to, primarily, vascular and infection, appears to have an intended effect of elevating missed stroke as an important area for concern in the Emergency Department. This is paired with further Newman-Toker citations regarding the frequency and harms of missed stroke. For example, Citation 110 (Newman-Toker DE 2016) is the reference for the claim of "45,000 to 75,000 missed strokes", which is actually an editorial. This editorial makes that same claim with regard to 45,000 to 75,000 missed strokes, citing as its source *another* Newman-Toker article (DOI: 10.1055/s-0035-1564298). This paper then cites as its source a Kerber study	The methods for this systematic review were described *a priori* in our protocol, with many processes in place to safeguard against bias. Inclusion and exclusion criteria were determined *a priori* and study authorship was not an inclusion or exclusion criterion. As a part of our processes to safeguard against bias individuals authoring studies were not involved in screening of studies for inclusion and exclusion. Finally, all authors of the report recused themselves from risk of bias assessment and data extraction of their own studies. We have added language to the front matter about these processes to manage potential conflicts of interest by the authors of the report.

Commentator & Affiliation	Section	Comment	Response
Ryan Radecki (cont'd)	methods/references (cont'd)	(10.1161/ 01.STR.0000240329.48263.0d) in which 1,666 patients with dizziness are evaluated, 46 of which are diagnosed with a stroke or TIA, and 16 of which were missed. The more appropriate estimate for the frequency with which dizziness is misdiagnosed should remain consistent with the denominator of this study, which is the ~4.4M emergency department visits annually in the U.S. for dizziness. Thus, including the remaining body of research using look-forward methods for subsequent stroke diagnoses in those discharged from the ED, the frequency of missed stroke in the ED is rather in the range of the 0.2% cited previously, with high-end estimates of around 0.5%, or 9,000 to 22,000 annually. This pattern of self-citation, in which further self-citation occurs, suggests the draft authors have unmanageable biases towards their own field of professional interest. In summary, this report should be rejected based on its many flaws. The framing and interpretation of the analyses performed does a disservice to the great unmet need for high-quality prospective data regarding diagnostic error in the Emergency Department.	(1) "elevating missed stroke as an important area for concern in the Emergency Department" --- The evidence base is quite clear --- stroke is the leading cause of serious misdiagnosis-related harms in the ED (particularly as it relates to malpractice claims, but bolstered by other data sources). The malpractice claims database from which these analyses derive (CRICO Comparative Benchmarking System) is a data source that is highly representative of US malpractice claims. The data set includes 20 member insurers that contribute their malpractice claims for coding and comparative analysis and together represent about 30% of all claims in the US; cases come from all 50 states plus Washington, D.C. and Puerto Rico, and the overall distribution of cases closely mirrors the findings from the National Practitioner Data Bank (see PMID: 31535832 [specifically Appendix A2]). A previous CRICO study from 2011 (not involving any of the current study authors) found the same --- fractures are the most common ED claim and stroke the second most (https://www.rmf.harvard.edu/~/media/Files/_Global/KC/PDFs/ crico_benchmarking_emergency_medicine.pdf), but when serious harms are considered, strokes far outnumber fractures (and outnumber all other causes). Other studies using different data sets have again found the same - for example, The Doctor's Company recently reported on ED misdiagnoses (Ross, 2021) stating, "The top categories for final diagnosis among the settled claims differed slightly. The highest classification remained cerebrovascular disease, but at a larger percentage (18 percent)." (https://www.thedoctors.com/siteassets/pdfs/ risk-management/closed-claims-studies/ emergency-department-process-of-care-closed-claims-study _focus-on-diagnosis-case-types.pdf) (we have added text to this effect to the report's Discussion to bolster this point).

Commentator & Affiliation	Section	Comment	Response
Ryan Radecki (cont'd)	methods/references (cont'd)	(comment above)	Lastly, for KQ2, the systematic review included 50 studies of missed stroke in the ED; the next most for KQ2 was myocardial infarction with just 15 studies --- it is hard to imagine why there would be so many more studies of missed stroke (by non-report authors) if this were not recognized as a known problem area in the ED; this point is further corroborated by the high rate of diagnostic error measured for stroke (17% false negative rate) relative to myocardial infarction (1-2% false negative rate). (2) "This is paired with further [report author] citations regarding the frequency and harms of missed stroke" --- The numbers cited by the commenter are roughly correct, although a more recent estimate of dizzy patients in the ED from NHAMCS is 4.8M rather than 4.4M. The 45,000-75,000 range is an estimate of diagnostic errors (i.e., missed dizzy strokes) and the 9,000-22,000 range is an estimate of misdiagnosis-related harms (i.e., stroke hospitalizations after missed dizzy strokes). These are entirely consistent with one another, as articulated in response to the prior comment by this commenter about choice of denominator. Errors will always outnumber harms from errors, since some patients whose dangerous diseases are missed "get lucky" and don't suffer any immediate or short-term consequences (this is true for stroke, but also all other conditions, though the rate of "getting lucky" varies substantially by disease).

Commentator & Affiliation	Section	Comment	Response
Mark Graber	Discussion	The discussion section of the report is problematic; it is highly biased, and generally weak. The main recommendation, to focus on disease-specific diagnosis, is reasonable as ONE possible approach to reducing harm from error in the ED, but it is hardly the only reasonable approach. Other approaches are equally promising, and at this very early stage of interventions to address diagnostic error, it would be a big mistake to put all the eggs in one basket. Although the authors argue that general solutions are unlikely to be effective, this conclusion is premature and not based on evidence. In fact there is substantial evidence that â€˜general solutionsâ€™ like education and training do correlate with better outcomes. Also, as part of an evidence-based review, it should be clearly noted that there are to date almost no large studies of the â€˜general solutionsâ€™ type; At this point in time, it is very inappropriate to conclude that â€œAll solutions wlll likely need to be tailored on a â€¦specific basis." Other avenues need to be mentioned and discussed: patient-based interventions, physician-focused interventions, and cognitive interventions in particular all need consideration."	Thank you. We agree that there may be multiple solutions to address the issue. While "solutions" or interventions are outside the scope of the report, we did offer some suggestions in the report for how our findings may influence thinking about the general direction or form future solutions might take (i.e., to seek to address the sorts of problems identified in the review as major problems, both in terms of priority diseases/symptoms and in terms of how demonstrated causes interact with solution sets). On the issue of disease-agnostic solutions, two systematic reviews (2013, 2019) on interventions to prevent diagnostic errors or mitigate harms (PMID: 23460094, 34408064) found just 2 of 129 solutions that were truly disease-agnostic (neither of which studied a patient outcome) and 4 others that addressed critical lab value alerts (2 of which showed no benefit and 2 of which didn't study outcomes). All of the studies showing benefit were disease-specific. Nevertheless, in response to the commenter's concern, we have revised the language in the Discussion section, changing "ample evidence" to "some evidence" ("Although this would seem to be the quickest way to solve the problem of ED diagnostic error, there is some evidence to suggest that general solutions like this are unlikely to work.").
Mark Graber	page 81	Minor point: Page 81. Implications for Clinical Practice, Education, Research, or Health Policy. In this section, the first sentence can be deleted; it isnâ€™t clear what barriers are being discussed	Thank you for pointing this out. During an earlier stage of responses to review, paragraphs were changed in location, and that sentence is now out of place, having become disconnected from the associated paragraph about barriers. It has been deleted.

Commentator & Affiliation	Section	Comment	Response
Mark Graber	Page 81	Second sentence: Eliminating diagnostic error is not a reasonable expectation; the field, and the text here, should focus on ways to minimize error or mitigate it.	We agree with this comment that eliminating all diagnostic errors or all associated harms (i.e., "zero harms") is probably not a realistic goal, so have added the word "preventable" each time in the report where we used the word "eliminate" or "eliminating" in reference to errors or harms so that it now reads "eliminating preventable misdiagnosis-related harms in the ED" or similar.
Mark Graber	Overall	Overall excellent report; the discussion of interventions is the obvious weak point; it presents a highly biased viewpoint; The many other ways to address diagnostic error (outside of the disease-focused ones) need to be presented with equal evidence, and discussed.	Thank you for the kind words on the overall report. Issues related to the Discussion section are addressed above in response to the commenter's first comment.
Denise Bockwoldt	Methods	The data was derived from malpractice claims, which is an example of structural and racial bias. Persons of color, poor persons, and immigrants are unlikely to file a malpractice claim. Thus, your data can only be generalized to those who have the means to acquire counsel. "In our analysis, the most comprehensive data on disease distribution among cases of diagnostic error were from large malpractice claims and incident report studies."	We agree that structural racism and racial bias are important problems, and likely to impact who will file a malpractice claim. The ED is often utilized by people of color and the economically disadvantaged, and these vulnerable populations (i.e., minorities/low SES/low health literacy) may be underrepresented in claims. We have added language to the section "Representativeness of Malpractice Claims Data for Disease Distribution" on this point ("Other biases could be at work that are not readily apparent from the available literature. For example, disadvantaged or vulnerable populations (e.g., those who are differently abled, racial or ethnic minorities, lower health literacy, lower socioeconomic status, prisoners) might be both more likely to be misdiagnosed and less likely to file a legal claim. However, we could find no specific evidence to suggest that this would likely impact the distribution of diseases for KQ1.").

Commentator & Affiliation	Section	Comment	Response
Denise Bockwoldt	Results	The data was derived from malpractice claims, which is an example of structural and racial bias. Persons of color, poor persons, and immigrants are unlikely to file a malpractice claim. Thus, your data can only be generalized to those who have the means to acquire counsel. "Malpractice claims are routinely captured by risk insurers, labeled as diagnostic error-related, and then thoroughly analyzed, making them a rich source of information on the distribution of diseases (KQ1) and causes (KQ3) of diagnostic error."	The issue of KQ1 and structural/racial bias in malpractice data is addressed in the prior comment. In addition, we have expanded the section describing gaps for KQ3 ("Other patient characteristics reflecting marginalized status as defined in AHRQ priority populations (e.g., members of religious minorities; lesbian, gay, bisexual, transgender, and queer [LGBTQ+] persons; persons with disabilities; persons who live in rural areas; and persons otherwise adversely affected by persistent poverty [including homelessness] or inequality) or marginalizing co-morbidities (e.g., mental health or substance use disorders) that may increase the risk of diagnostic error are understudied and deserve further equity-related research.")
Denise Bockwoldt	General	I believe the research question should be on "WHO is harmed" rather than "WHAT harm took place. Marginalized patients, such as the obese, persons of color, female gender, the poor, and non-English speaking are at high risk of misdiagnosis due to implicit biases. Data on diagnostic errors should be reported with demographic data as well.	Some studies have clearly demonstrated that minority and marginalized patients are at higher risk of misdiagnosis, and this is addressed in the "Patient Characteristics" section of KQ3. We did not find studies that proved implicit bias was the cause of these disparities in diagnosis, although we agree this may be one underlying root cause of the demonstrated disparities. Demographic data are unavailable for the large malpractice and incident report studies used to address KQ1 and part of KQ3. As noted in response to the prior two comments, we have identified this issue as a gap that should be filled through further equity-related research in diagnosis. We have also added language to clarify the importance of this issue ("Achieving equity in diagnosis by addressing racial and other diagnostic health disparities is of recognized importance to achieving diagnostic excellence."; "The root causes of measured diagnostic disparities should be examined, including the role of implicit or explicit bias towards women, minorities, or other vulnerable populations."; "To summarize, measuring health equity in diagnosis should be a key focus of future research, and special care should be taken to ensure that rigorous epidemiologic and statistical methods are used to address this concern, since incorrect methods can lead to erroneous inferences.")

Commentator & Affiliation	Section	Comment	Response
Timothy Hofer	Evidence Summary	1/ This report says little about the problem of measurement of diagnostic errors nor does it make any reference I can find to such a problem. In other words the problem of how we determine that an error is present and how reliably we are able to determine that it is present. Most expert judgement has low reliability for determining the presence or absence of a preventable error with the best quality studies of measurement characteristics finding reliabilities between 0.2 and 0.6.	Thank you for pointing this out. We have added to the Executive Summary (and report Conclusions) the following statement: "Our review findings are tempered by limitations in the underlying evidence base, including issues linked to data sources, measurement methods, and causal relationships." We have also added a section to the Limitations to address measurement concerns and issues of bias in determination of errors expressed by another public commenter: "Most studies did not directly address issues surrounding measurement of diagnostic error (e.g., validity, reliability, determination of causes, preventability, or attribution of harms). In clinical practice, many disease reference standards are insufficiently understood, developed, and implemented, so diagnosticians often disagree on final patient diagnoses. To the extent that manual chart reviews were used to identify errors, original studies are likely to suffer from problems of poor chart documentation, low inter-rater reliability, and hindsight bias. The problem of author bias in choice of definition or method of measurement (e.g., specialists [or diagnostic error “advocates”] determining ED misdiagnosis and favoring more lax definitions of error/harm, or the reverse, with ED clinicians favoring more stringent definitions) is difficult to ascertain. Our use of the NAM definition of diagnostic error mitigates some of these concerns, since there is less subjectivity inherent in a diagnostic label change (e.g., discharged with “musculoskeletal chest pain” returns with “aortic dissection” within 24 hours) than in the determination of preventability, which is known to be highly subjective. Also, many included studies used stringent measurement protocols or objective statistical methods (e.g., SPADE). Nevertheless, poorly standardized or low-reliability measurements are important limitations."

Commentator & Affiliation	Section	Comment	Response
Timothy Hofer	Evidence Summary	2/ Malpractice claims are very problematic as a source as was clearly shown as far back by the Harvard Medical Practice Study where in a review of 50,000 records there was exactly no overlap between expert judgement about when an error occurred and a malpractice claim determination of an error. It seems like it might be worth mentioning this as a limitation.	Thank you for this comment. We have added to the Executive Summary (and report Conclusions) the following statement: "Our review findings are tempered by limitations in the underlying evidence base, including issues linked to data sources, measurement methods, and causal relationships." While we agree there are problems inherent in the use of malpractice claims data, the stated lack of overlap between an expert determination of "error" and malpractice claim determination of error is not necessarily germane to the report's findings. Importantly, we believe the commenter intends that the term "error" (when determined by an expert) indicates that a process failure occurred, which diverges from the NAM definition of diagnostic error used in this report. We offer extensive treatment of the issues surrounding malpractice claims in the body of the report (for example, see "Representativeness of Malpractice Claims Data for Disease Distribution" and "Representativeness of Malpractice Claims Data for Root Causes").

Commentator & Affiliation	Section	Comment	Response
Timothy Hofer	Evidence Summary	3/ Perhaps most importantly, the report avoids talking about the very difficult step of going from saying that a process labeled as an error is present and a bad outcome has occurred to saying that an error has caused the bad outcome and attributing the outcome entirely to the error. This should be presented as a major limitation of any attempt to quantify the impact of diagnostic errors in the emergency department."	We appreciate this distinction. We have added to the Executive Summary (and report Conclusions) the following statement: "Our review findings are tempered by limitations in the underlying evidence base, including issues linked to data sources, measurement methods, and causal relationships." In the Strengths and Limitations section, we have added text describing limitations further, including this issue of preventability or attribution of harms ("Most studies did not directly address issues surrounding measurement of diagnostic error (e.g., validity, reliability, determination of causes, preventability, or attribution of harms). In clinical practice, many disease reference standards are insufficiently understood, developed, and implemented, so diagnosticians often disagree on final patient diagnoses. To the extent that manual chart reviews were used to identify errors, original studies are likely to suffer from problems of poor chart documentation, low inter-rater reliability, and hindsight bias. The problem of author bias in choice of definition or method of measurement (e.g., specialists [or diagnostic error “advocates”] determining ED misdiagnosis and favoring more lax definitions of error/harm, or the reverse, with ED clinicians favoring more stringent definitions) is difficult to ascertain. Our use of the NAM definition of diagnostic error mitigates some of these concerns, since there is less subjectivity inherent in a diagnostic label change (e.g., discharged with “musculoskeletal chest pain” returns with “aortic dissection” within 24 hours) than in the determination of preventability, which is known to be highly subjective. Also, many included studies used stringent measurement protocols or objective statistical methods (e.g., SPADE). Nevertheless, poorly standardized or low-reliability measurements are important limitations.").

Commentator & Affiliation	Section	Comment	Response
Timothy Hofer	Evidence Summary	The medical literature has a predilection for substantially overstating causal claims based on weak designs and evidence and this report is a some risk of being a prime exemplar. While I admire the enthusiasm of the report writers, I would suggest that the tone of certainty be modulated and that conclusions and recommendations be made with a greater spirit of humility.	We have added to the Executive Summary (and report Conclusions) the following statement: "Our review findings are tempered by limitations in the underlying evidence base, including issues linked to data sources, measurement methods, and causal relationships."

Commentator & Affiliation	Section	Comment	Response
Charles A Pilcher	Conclusion	This report overlooks the power of stories as teaching tools. Emergency physicians are taught too little about "pitfalls": What will you miss? Why will you miss it? How do you avoid missing it? The solution is simply telling stories. If only one story for each of the 9 most missed diagnoses were told to each medical student and EM resident, that story will be remembered better than any of the solutions this paper proposes. [See (1) Aronson L. Story as Evidence, Evidence as Story. JAMA. 2015;324(2):125-6. https://chcimedicalhumanities.org/media/cms_page_media/25/BiblioSummerInstitute-Aronson_JAMA_2015.pdf and (2) Boris V. What Makes Storytelling So Effective For Learning? Harvard Business Publishing: Corporate Learning. December 20, 2017. Accessed at https://www.harvardbusiness.org/what-makes-storytelling-so-effective-for-learning/] The paper speaks of "modular learning" and acknowledges that "general solutions... are unlikely to work." The paper further suggests that since "cognitive errors in diagnostic reasoning predominate... solutions will likely need to be tailored on a symptom- and disease-specific basis (i.e., modular). Stories are the simplest way to do that and have the learnings remembered. We begin to learn from stories in infancy. They are the most "modular" form of learning for the human race.	We concur that stories are often quite powerful and could be used to highlight common pitfalls in diagnosis (PMID: 35061037). Providing or compiling such stories in the form of clinical examples is beyond the scope of this systematic review. However, many such stories have already been assembled by the Society to Improve Diagnosis in Medicine (https://www.improvediagnosis.org/stories/) and other patient safety-focused organizations.

Commentator & Affiliation	Section	Comment	Response
Charles A Pilcher	Conclusion	The paper acknowledges that EP's rarely get performance feedback.	We concur that stories are often quite powerful. We understand how telling a story about an egregious mistake in a patient whose dizziness was due to stroke but was misattributed to inner ear disease and led to devastating patient harms (e.g., a young athlete being left quadriplegic and mute... the so-called "locked in syndrome" [https://www.improvediagnosis.org/stories_posts/missed-stroke-diagnosis/]) would be quite memorable and frightening to an ED clinician. What is unclear to us is how such a story would help the average ED physician correctly perform and interpret bedside eye movement maneuvers that might have prevented such a mistake --- ED physicians are already quite worried about missing stroke in dizziness, yet have important gaps and misconceptions in how to evaluate patients with dizziness (PMID: 26231272) and remain uncomfortable with the bedside skills required to differentiate stroke from inner ear disease (PMID: 26587108).

Commentator & Affiliation	Section	Comment	Response
Charles A Pilcher (cont'd)	Conclusion (cont'd)	There's an indirect way of providing that feedback by accessing the learnings from our most egregious mistakes, the ones that become med mal lawsuits. The least defensible of those are settled pre-trial. The stories of these lawsuits are then buried in non-disclosure agreements, yet they are the cases with the greatest likelihood of helping a budding EP never make the same mistake that a colleague made. Reading or hearing a brief story of a case gone wrong is far more effective and memorable than a 1 hour lecture on the topic. Telling a mere 9 stories for the "target diseases" that the article prioritizes will reduce the rate of "misdiagnosis-related harms (particularly high-severity harms)." The transparency that a story bring also honors the injured patient (posterior circulation stroke, aortic dissection, spinal epidural abscess, necrotizing fasciitis, etc.) whose reason for suing always includes the wish that the same thing would never happen to someone else. It also is an answer to the complaint that educators have "too much to teach and too little time to teach it." As an example, the article promotes the construction of "clinical practice guideline for acute dizziness" to identify strokes. A simple story of a missed diagnosis that became a lawsuit will do the same thing a lot faster and far more memorably. A busy EP or EM resident will benefit more from that story than spending his/her time reading a guideline "currently under development by the Society for Academic Emergency Medicine.299"	(response above)

Commentator & Affiliation	Section	Comment	Response
Charles A Pilcher	Conclusion	Finally, the paper notes that there is "...downward financial pressure on use of MRI in back pain presentations [that] may increase the risk of missed spinal abscess, which requires spine imaging for diagnosis." While true, there are 3 steps available to EP's that must be taken before even considering an MRI in a patient with spine pain. The first is simply asking oneself "Could this patient have a spinal epidural abscess (SEA)?" A diagnosis never considered will never be made. The second is the history. Once considered, SEA can be ruled out by history alone in 95% of patients. This assumes the absence of the "classic trial" for SEA of back pain, fever and neurological symptoms, which is present in less than 20% of SEA patients. (reference available). Third, if SEA remains a possibility after history and exam, a CRP and/or ESR should be obtained. The test will be grossly elevated in over 95% of patients (references available) and, if normal, allows the patient to be treated symptomatically, followed closely and an MRI obtained only if symptoms progress.	The statement about downward financial pressure related to spinal abscess was added at the suggestion of a prior reviewer. We concur that failures of bedside history taking and examination represent a central issue in misdiagnosis, as identified in KQ3. To make this point more directly and forcefully, we have modified text in the section on "Considerations for Clinical Practice and Policy". ("But this “tradeoff” scenario assumes that (a) current practice optimally applies existing diagnostic methods, (b) innovations in diagnosis do not occur, and (c) that as a consequence, the only way to influence diagnosis is to alter the threshold for ordering existing tests (e.g., by lowering the threshold and testing patients at very low risk for the target disease). In turn, this premise leads to the (often) erroneous conclusion that diagnosis is a “zero sum game” and the only choice is to “pick your poison” between more false negatives (favor specificity, sacrifice sensitivity) and more false positives (favor sensitivity, sacrifice specificity). However, this is generally a false dichotomy, since current practice often fails to apply basic diagnostic methods (e.g., proper history-taking and neurologic examination in patients with back pain at risk for spinal abscess [Bhise, 2017]) and innovations that actually improve diagnosis (e.g., via better education/training, new clinical pathways, novel diagnostic tests, enhanced teamwork in diagnosis, greater access to specialists, or improved feedback/calibration) will almost always increase both sensitivity and specificity at any given decision threshold. The result is then fewer false negatives and fewer false positives, sometimes even at a lower total cost.")

Commentator & Affiliation	Section	Comment	Response
Charles A Pilcher	Conclusion	Dr. Mark Graber and I have a paper currently in review that focuses on the power of stories. The transparency that we advocate is similar to the way other industries address mistakes, examples being OSHA, CPSC, NHTSA and NTSB. If the aviation industry managed its plane crashes like we in healthcare manage ours, none of us would dare get on an airplane. And if we continue to bury the learnings from our mistakes that become lawsuits, we will continue to bury our patients. Thanks for allowing me to comment - or rant. I'm just a retired pit doc who has seen too many avoidable mistakes made because we haven't been able to learn from the mistakes of others. I wish this project well.	We concur that stories are often quite powerful. We are glad you are bringing this issue to a broader audience with Dr. Graber.
Charles A Pilcher	General	It's too long, too pontificating and takes such a high level approach that it will take decades to become standard education for the next generation of emergency physicians. I'm an advocate for stories. They work better and faster than academic treatises. I also believe that risk managers, med mal insurers and attorneys must acknowledge that they are best positioned to improve patient safety by allowing, encouraging or even requiring greater transparency in the outcomes of our worst mistakes, particularly those that are settled without trial for over $1 million.	We concur that stories are often more memorable and compelling than detailed reports. However, the Task Order from AHRQ was for a systematic review of available evidence on the topic of diagnostic errors in the ED.

Commentator & Affiliation	Section	Comment	Response
Paula Distabile	Discussion	This comment is re: Report page 86 - pdf page 98 â€“ paragraph 2.b.: "The AHRQ Common Formats for Event Reporting (CFER) now include a special common format for Diagnostic Safety event reporting (CFER-DS) that has recently been developed for use by patient safety organizations (PSOs)." We very much appreciate mention of the CFER-DS. A small addition may be helpful here to clarify that the CFER-DS (and all of the other AHRQ Common Formats) are available in the public domain to encourage their widespread adoption. An entity does not need to be listed as a PSO or working with one to use the Common Formats. However, it should also be noted that the Federal privilege and confidentiality protections only apply to information developed as patient safety work product by providers and federally listed PSOs working under the Patient Safety and Quality Improvement Act of 2005.	Thank you for the helpful comment. We have added this information to the relevant section ("The CFER-DS (and all of the other AHRQ Common Formats) are available in the public domain to encourage their widespread adoption. An entity does not need to be listed as a PSO or working with one to use the Common Formats. However, it should also be noted that the Federal privilege and confidentiality protections only apply to information developed as patient safety work product by providers and federally listed PSOs working under the Patient Safety and Quality Improvement Act of 2005.").

Commentator & Affiliation	Section	Comment	Response
Anonymous	Methods	The most serious limitation of this review is that is seems to be biased in many ways towards identifying those diseases and using those data sources that are of interest to the authors. â€¢ First, the main research question of this manuscript is â€˜What diseases or syndromes are associated with the greatest total number and the highest risk of diagnostic errors or misdiagnosis related harms?â€™ Surprisingly, the authors seem to have decided on what those diseases are before conducting the actual review, as they excluded diseases that they did not consider relevant. Based on a search in the Appendix, it seems that 296 articles were excluded because the â€˜Population does not have the condition of interestâ€™, which is an incredibly large number. This review therefore does not answer the proposed research question, but predefined a set of diseases that the authors consider impactful and examined only studies with those diseases involved. Based on the current approach, question 1 is therefore not answered.	The list of pre-specified diseases was not used to answer KQ1 (i.e., what disease or syndromes are associated with the greatest number and highest risk of errors?), but to answer KQ2 (“Overall and for the clinical conditions of interest, how frequent are ED diagnostic errors and associated harms?”). For KQ1, none of the disease-specific studies was used, as this would not permit frequency comparisons across diseases. As described in the Methods, the list of diseases under consideration was formulated on the basis of prior literature and informed by input from a Technical Expert Panel and Key Informants. Now having analyzed the results from KQ1, the prespecified list appears to have been fairly complete vis-a-vis the most common causes of misdiagnosis-related harms - for example, in the largest incident report study of ED diagnostic errors (n=2,288) (which was not used to determine the prespecified list), all top 12 conditions found in that study (see Table 1 from PMID: 31801474) appeared in our prespecified list. No other conditions identified in that study had higher individual frequency, and, collectively, those "other" conditions outside the top 12 accounted for just 30% of the total incidents reported (n=679/2,288). While some conditions (particularly those affecting children) may have been underrepresented (e.g., missed child abuse/non-accidental trauma), we found no evidence to suggest that using a prespecified list based on prior literature, and input from our Technical Expert Panel and Key Informants, appreciably affected the overall results. We have added a paragraph to the Limitations section describing this issue.

Commentator & Affiliation	Section	Comment	Response
Anonymous	Methods	Secondly, studies with fewer than 50 cases were excluded. While it makes sense to exclude case reports, there is no reason to exclude studies with a smaller sample size. In prospective studies, or studies including consecutive patients, the number of 50 error cases is actually quite high. Excluding those studies from the review will cause bias, as those smaller studies use a method/data source which can identify different types of diseases than studies with large databases. The strength of a systematic review is that it pools different type of studies, with different methods. The authors mention themselves in the review that â€˜It is usually necessary to rely on multiple data sources and different methods to gain a more comprehensive view of patient safety and quality.â€™ Excluding 87 studies because they have fewer than 50 patients is bias is not correct in a systematic review	Decisions about eligibility criteria were made *a priori* and documented in our protocol. The protocol was developed based on feedback from our Technical Expert Panel and Key Informants. The vast majority of studies had sample sizes between 50 and 4,000. Only 10% (n=120) of 1,176 excluded studies had a sample size <50 and only 5% (n=60) had this as the only listed exclusionary criterion. Of the 60 smaller studies, 24 had a sample size <10 and 18 of these had a sample size <5. Thus, we do not believe that this decision had a meaningful impact on the results, but, nevertheless, we have added a sentence to the Strengths and Limitations section listing this as a potential limitation ("We also do not know whether exclusion of smaller studies (n<50) by design influenced results.").

Commentator & Affiliation	Section	Comment	Response
Anonymous	Methods	The data sources reflected in this review rely largely on malpractice claims (this has the diseases that the authors are interested in, and have a large number of cases). While those reflect cases that harmed patients, they also reflect a subset of cases. Problems with malpractice claims (and incident reports) is that they include a biased sample which should not be used to measure prevalence of diagnostic errors. In this review they are used to make statements about prevalence and even the different sources are compared. In addition, the review relies heavily on one large malpractice claim database which is a serious weakness of the review.	(1) "data sources ... rely largely on malpractice claims" --- Only a handful of the 279 studies included in the systematic review are malpractice claims studies, and they are principally used to answer KQ1, with a smaller contribution to root cause determinations in KQ3. Malpractice claims data were not chosen because they have "diseases that the authors are interested in" but because they, alongside incident reports, represent the only meaningful source of consistently derived and coded data on the relative frequency of diseases associated with harmful diagnostic errors in the ED. (2) "they include a biased sample which should not be used to measure prevalence of diagnostic error" --- We agree, and prevalence questions in KQ2 were not answered using any malpractice data. Issues of representativeness of malpractice claims data for KQ1 (frequency distribution of diseases causing serious harms, NOT prevalence) are addressed in "Representativeness of Malpractice Claims Data for Disease Distribution" and for KQ3 are addressed in "Representativeness of Malpractice Claims Data for Root Causes".

Commentator & Affiliation	Section	Comment	Response
Anonymous (cont'd)	Methods (cont'd)	(comment above)	(3) "the review relies heavily on one large malpractice claim database" --- the specific answer to KQ1 relies heavily on two data sources --- one a large malpractice claims database and the other a large incident report analysis, whose findings were very closely matched on the list of diseases (i.e., the answer to KQ1). These two data sources combined included 4,561 cases, with the three next largest data sources relevant to KQ1 (all malpractice claims studies) having a total of just 357 combined cases coded in non-comparable fashion (making synthesis impossible). More importantly, the malpractice claims database from which these analyses derive (CRICO Comparative Benchmarking System) is a data source that includes 20 member insurers that contribute their malpractice claims for coding and comparative analysis and represents about 30% of all claims in the US; cases come from all 50 states plus Washington, D.C. and Puerto Rico, and the overall distribution of cases closely mirrors the findings from the National Practitioner Data Bank (see PMID: 31535832 [specifically Appendix A2]). Biases inherent in malpractice claims (e.g., towards more severe cases) mean we must consider the representativeness of malpractice claims for ED error and harms (see "Representativeness of Malpractice Claims Data for Disease Distribution"), but there is no evidence that the single claims study relied upon is biased with respect to all US malpractice claims; instead, there is strong evidence that it is far more representative of all US malpractice claims than any of the smaller studies. The fact that these two large studies represent 78% of the total meta-analytic sample was noted in the text ("There were nine studies that addressed KQ1a directly for all diagnostic errors, reporting on a total of 5,817 diagnostic errors. Four studies were malpractice claims-based and five were incident report-based. The two largest studies, one a large, United States-based review of a national malpractice claims database (Newman-Toker, 2019) and the other a large, United Kingdom-based review of a national incident reporting system (Hussain, 2019) together represented 78 percent of diagnostic error cases (n=4,561 of 5,817). These two studies organized their categories in similar enough fashion to present results together (Table 2).").

Commentator & Affiliation	Section	Comment	Response
Anonymous	Methods	Both malpractice claims and incident reporting systems are not set up with the aim of conducting scientific research. I have worked with many different malpractice claims databases and incident reporting systems, and the data are often of poor quality. Important quality indicators are often lacking (e.g. interrater reliability) and the quality of the questions used to analyze the claims are not validated. The same is true for incident reports. The review heavily relies on those data sources and they seem to have excluded studies with better research quality (with diseases not fitting the author scope and with better research quality (with diseases not fitting the authors scope and with [less-than] 50 cases).	We concur that malpractice claims and incident reporting systems are imperfect data sources and cannot be used to answer certain questions (like error or harm rates, which were answered in our report [KQ2] using non-claims data sources). However, for the purposes of identifying the most commonly identified diseases (i.e., the main answer to KQ1) associated with diagnostic errors, particularly those responsible for serious harms to patients, there is no better data source available than malpractice claims or incident reports. The commenter has not provided examples to support the contention that "studies with better research quality" were excluded, so we cannot respond. As to excluding studies with "diseases not fitting the author scope," this is incorrect --- for the purposes of KQ1, as described in response to another of this commenter's concerns, there was no pre-specified disease restriction --- in other words, there was no exclusion of studies on the basis of disease in the places where malpractice or incident reports were used. To make this point more clearly, we have modified the text in the Methods section ("As noted above in Data Sources, we searched for a mix of disease-agnostic (any disease) and disease-specific (one or a few specific diseases – e.g., major cardiovascular events) studies. The search strategy was designed to capture both sorts of studies. However, disease-specific studies could only be identified by pre-specifying these diseases as part of the search strategy.

Commentator & Affiliation	Section	Comment	Response
Anonymous (cont'd)	Methods (cont'd)		Based on preliminary knowledge of the literature, we proposed *a priori* the following conditions to be included in the disease-specific component of the search strategy: stroke, myocardial infarction, venous thromboembolism, aortic aneurysm and dissection, arterial thromboembolism, sepsis, meningitis and encephalitis, spinal abscess, pneumonia, endocarditis, appendicitis, and selected fractures. Additional conditions were added to expand the search based on input from Key Informants and the TEP. Additional conditions that are relevant to pediatric populations include testicular torsion, necrotizing enterocolitis, and sudden cardiac death/arrythmias/congenital heart disease. Additional conditions that are relevant to pregnant populations are ectopic pregnancy and preeclampsia/eclampsia. While screening the full-text articles, we included all disease-agnostic studies meeting our other entry criteria, but we excluded disease-specific studies that did not include populations with at least one of these named conditions. We did not exclude studies based on condition when screening abstracts."). We have also added a Table to the Appendix (Appendix A-4) linking specific KQs to particular data sources/types.

Commentator & Affiliation	Section	Comment	Response
Anonymous	Results	1.Fractures are considered to be overrepresented by the authors because of the easier final judgement that can be made with imaging. However, this is an assumption that is likely not true. It is often the diseases like stroke that are overrepresented in studies because of hindsight bias. In the many malpractice claims and cases that I reviewed, it was often the diseases like stroke in which an earlier ED visit with headache was considered a missed opportunity for diagnosing stroke. The fact that a stroke was not diagnosed in the first visit was often attributed to a diagnostic error was assumes a causal relationship while this was not necessarily there (it could be that a stroke was not there or not diagnosable yet). I would argue that for a fracture, it can be checked whether there was a diagnostic error, but for stroke it is not and often assumed and therefore overrepresented.	We apologize for the lack of clarity in our statement. First, we did not make this statement with certainty ("It is unknown whether ascertainment and reporting biases linked to radiographic misdiagnosis (which is more easily confirmed and contested than other types of diagnostic error) lead to fractures being further overrepresented in malpractice claims or incident reports, but their high annual incidence (2 million cases per year in the United States, as of 2020, according to the National Electronic Injury Surveillance System [NEISS]) makes it likely that, even if overrepresented, they are still quite common.").

Commentator & Affiliation	Section	Comment	Response
Anonymous (cont'd)	Results (cont'd)	(comment above)	Second, what we mean is that fractures are overrepresented relative to other less harmful diagnostic errors (not overrepresented relative to dangerous conditions, which are themselves overrepresented in malpractice claims and incident reports, as described in "Representativeness of Malpractice Claims Data for Disease Distribution"). For example, benign positional vertigo leads to nearly 1 million ED visits per year in the US and misdiagnosis rates are about 80-90% (see PMID: 21676060 and AVERT Trial Abstract presented at the International Barany Society, May, 2022 [first author Badihian, S.]). With 2 million fractures per year in the US (from the National Electronic Injury Surveillance System), the diagnostic error rate for ALL fractures would have to be at least 40% to exceed BPPV misdiagnoses, while the estimated rate from our systematic review suggests it is about 1%. Conversely, if there are 20,000 missed fractures (2M x 1%), then the error rate in diagnosing benign positional vertigo would have to be 2% for there to be the same number of BPPV misdiagnoses... the true diagnostic error rate is 91% (in a prospective, randomized trial with masked outcome determination) and the absolute, preventable diagnostic error rate is 60% (again, using a prospective, randomized trial design - AVERT trial NCT02483429). Thus, the estimated number of diagnostic errors with BPPV is 900,000; preventable errors is 600,000 --- while for fractures it is 20,000. Yet BPPV appears nowhere in any incident reports or malpractice claims, despite being 30- to 45-fold more common. This is what we mean when we say that fractures are potentially overrepresented (we have added text to KQ1a to elaborate on this issue). As to the issue of missed strokes, with a miss rate of 17% and an annual incidence of ~800,000 for completed strokes and about ~400,000 TIAs per year (PMID: 35078371), that would be ~200,000 missed strokes --- roughly ten times as many as missed fractures in the US. We have modified the text in multiple places to elaborate on this issue (by incorporating the arguments contained in this response), to avoid confusion.

Commentator & Affiliation	Section	Comment	Response
Anonymous	Discussion	1.In the recommendations it is suggested that research on reducing diagnostic errors in the ED should focus on the clinical presentations of the diseases in the top 10 rather than on the diseases itself. This makes sense since a patient presents with a symptom and not with a disease. However, the cases represented in review are largely based on malpractice claims, where diseases present atypically. This would then include a very large range of symptoms, which makes me wonder why we need to specify in the first place. Isnt it better to focus our efforts on common causes of errors than on specific diseases/ disease presentations?	(1) "the cases represented in review are largely based on malpractice claims" --- We do not agree with the premise of the question. While the list of diseases in question derives from malpractice and incident reports (KQ1a), all of the other data relevant to the issue of atypical symptoms are derived from non-claims, non-incident report sources, including prospective studies (see KQ3d and new Appendix on Data Types and Sources). It is certainly true that while there might be only 1-3 "classic" symptoms of a disease and two dozen or more "atypical" symptoms of a disease, it turns out that these are not evenly distributed randomly across the two dozen atypical symptoms --- instead there are "typical" atypical symptoms that are recurrent pitfalls. This is shown epidemiologically using SPADE look-back methods that help identify patterns of misdiagnosis where observed rates can be compared to expected rates for the frequency of these symptoms (e.g., PMID: 32701479, 33650389, 28344918). In such studies, there are typically only a handful of antecedent symptoms (i.e., "typical" atypical cases or "recurrent pitfalls") that account for the majority of missed cases. Thus, the problem is not intractable. We have modified the Discussion section to reflect this argument ("This suggests that system-wide, scalable solutions need to be developed to tackle cognitive problems, and that these solution sets must be targeted to address not the most common clinical presentations of key diseases of interest but the most commonly misdiagnosed clinical presentations of key diseases of interest. This is a tractable approach because epidemiologic studies using the SPADE look-back method have shown that only a handful of symptoms account for the majority of missed clinical presentations for any one disease—in other words, these are what might be called “typical” atypical cases or recurrent diagnostic pitfalls.") (2) "Isn't it better to focus our efforts on common causes of errors than on specific diseases/ disease presentations?" Since the systematic review found that the most common causes of diagnostic errors are cognitive issues of clinical expertise and bedside reasoning in atypical cases, this is precisely what focusing on recurrent pitfalls in diagnosis (which, by their nature are disease and symptom-specific) is striving to do.

Commentator & Affiliation	Section	Comment	Response
Anonymous	Discussion	2. Risks of overdiagnosis are not considered. The malpractice claims databases that this review largely builds on represent cases that present atypically or with less severe symptoms. To timely diagnose those cases, diagnostic testing has to be done for patients with a relatively low risk of a serious diseases. If this were to be implemented focused on the diseases and disease presentations identified in this review, this could result in serious overdiagnosis. Not every patient with a headache should get an MRI. Especially with measures such as SPADE, this is a risk. I think it is a nice method for benchmarking. However, even if a patient the non-specific headache or dizziness presented to the ER a week before a stroke was diagnosed, this does not mean that all of those patients had a diagnosable stroke the week before.	(1) "Risks of overdiagnosis are not considered" --- Overdiagnosis does occur in the ED (PMID: 35249191), but it appears the thrust of this comment is about the risks of diagnostic test overuse (rather than overdiagnosis, per se) (PMID: 29367314). It is a legitimate fear that a focus on improving sensitivity will sacrifice specificity. Issues surrounding test overuse and underuse are addressed in the report in the Section on Considerations for Clinical Practice and Policy (see lengthy paragraph that begins "In considering implications for clinical practice and policy, it is important to examine the apparent tension between test underuse and test overuse as it relates to diagnostic errors...") We have now added text about overdiagnosis to this section in response to this comment ("Furthermore, ED overuse of increasingly sensitive diagnostic tests now risks overdiagnosis of mild forms of illness where, despite a correct diagnosis, harms (physical, psychological, or financial) ultimately outweigh treatment benefits (e.g., sub-segmental pulmonary embolism).").

Commentator & Affiliation	Section	Comment	Response
Anonymous (cont'd)	Discussion (cont'd)	(comment above)	(2) "...with measures such as SPADE, this is a risk... this does not mean that all of those patients [with headache and discharged] had a diagnosable stroke the week before" --- In this comment, there are two separate issues: (a) biological/causal relationship---i.e., is the prior headache antecedent to the stroke *related* [a marker of the evolving disease process] or *unrelated* [coincidental] to the stroke that happens a week later?; and (b) preventability of the error---i.e., was the headache "diagnosable" as a stroke manifestation, with a presumption built in about "reasonableness" or "appropriateness" of efforts that might be required to diagnose the stroke (number needed to diagnose, cost-effectiveness of diagnosis, etc.). On the former, the SPADE method uses observed minus expected approaches to determine the presence of a biological/causal relationship (i.e., non-coincidental relationship) (PMID: 32701479); this is far stronger than any chart review-based method and has been shown to be statistically robust (PMID: 34115418). On the latter, preventability at reasonable cost is an important issue for determining appropriateness of solutions. However, the current state of the science on preventing diagnostic error is still in its infancy. Thus, using "cost-effectiveness of solutions" (that do not yet exist) to determine what diagnostic errors we choose to measure (or not) is premature. Instead, we should measure diagnostic errors as robustly as we can and seek to solve problems that can be solved at reasonable cost. As to the current "preventability" of stroke misdiagnosis in patients with headache, the results, at least for the subset with missed subarachnoid hemorrhage (SAH) (PMID: 30797572), are fairly clear---it is quite likely that proper widespread application of validated bedside clinical decision rules (PMID: 24065011) could reduce missed SAH at *lower* cost (PMID: 31805846).

Commentator & Affiliation	Section	Comment	Response
Anonymous	Discussion	3. The recommendations do not logically flow from the findings of the review. While cognitive errors are the most common, there is limited attention for those causes in the recommendations. In general, the main causes underlying diagnostic errors should get more attention.	(1) We believe that the suggested policy considerations flow directly from the report's findings: "(1) standardizing measurement and research results reporting to maximize comparability of measures of diagnostic error and misdiagnosis-related harms" - this derives directly from the lack of standardized measurement of diagnostic error and harms identified by the systematic review; "(2) creating a National Diagnostic Performance Dashboard to track performance (analogous to the Dartmouth Atlas Project for utilization of healthcare services)" - this derives from the lack of adequate national benchmarking and lack of comparability of measurement across EDs identified in this systematic review; and "(3) using multiple policy levers (e.g., research funding, public accountability, payment reforms)1 to push for the rapid development and deployment of solutions that address this major patient safety and quality problem" - this derives directly from the overall public health scale/scope of the problem identified by the review.

Commentator & Affiliation	Section	Comment	Response
Anonymous (cont'd)	Discussion (cont'd)	(comment above)	(2) "While cognitive errors are the most common, there is limited attention for those causes in the recommendations" --- While "solutions" are outside the scope of the report's overall charge, we did offer some suggestions for how our findings may influence thinking about solutions (i.e., that address the sorts of problems identified in the review, both in terms of priority diseases/symptoms and in terms of how demonstrated causes interact with solution sets). We agree the most common errors are cognitive (and so state) but disagree that our policy considerations fail to address these. As we noted above in reply to an earlier comment from this individual ("Isn't it better to focus our efforts on common causes of errors than on specific diseases/ disease presentations?") --- since the systematic review found that the most common causes of diagnostic errors are cognitive issues of clinical expertise and bedside reasoning in atypical cases, this is precisely what focusing on recurrent pitfalls in diagnosis (which, by their nature are disease and symptom-specific) is striving to do. We did not propose solutions to this problem in other than the general sense of pointing to teamwork, training, and technology as likely mechanisms to address cognitive errors and the need for modularity in addressing them. The relevant text is in the Discussion section focused on KQ3 ("Taken together, this suggests that interventions to reduce harm from ED diagnostic error must directly tackle problems in fundamental bedside diagnostic skills and clinical reasoning for atypical presentations of the ten diseases producing the most harm. If substantial headway is to be made, we must develop system-wide solutions to address these cognitive problems. Options fall into three basic mechanisms that all target increasing the availability of diagnostic expertise: (1) build the expertise of ED clinicians through deliberate practice training and feedback; (2) support ED clinicians' decision-making through teamwork, including access to experts; (3) minimize cognitive load by deploying technologies that digitally encapsulate expertise. Because diagnostic expertise is deeply problem-specific, these broadly construed solutions will need to be individually tailored on a symptom- and disease-specific basis (i.e., modular).")

Commentator & Affiliation	Section	Comment	Response
Anonymous	Conclusion	The conclusions cannot be made based on on the method that was applied. Too many studies that are potentially relevant van been excluded and the authors. I therefore think that the conclusion are not correct	The methods were determined *a priori* with input from our Technical Expert Panel and Key Informants. As noted in other responses, we disagree with the assertion that relevant studies were excluded. We followed our protocol and have acknowledged the limitations in the evidence based and in our methods.
Anonymous	References	From the references it is clear where the authors are from as there is a lot of self-citation. Other groups with a lot of work on diagnostic error are barely cited. For example, there is a part in the review of electronic trigger tools where the Houston group has done much work but is barely cited (e.g. paper by Dr Murphy).	All articles for the systematic review were assessed with respect to the inclusion and exclusion criteria defined *a priori* in our protocol. Study authors were not involved in the screening of their own studies for inclusion in the systematic review. The work by the Houston group on electronic triggers is focused largely on cancer diagnosis in primary care. It is therefore not surprising that their work would be underrepresented in the current report about ED diagnostic error. The commenter has not specified the "paper by Dr. Murphy," and we were not able to locate a specific study by Dr. Murphy regarding electronic triggers in the ED; another of the Houston group's electronic trigger studies relevant to the ED is included in the report (PMID: 34279630).

Commentator & Affiliation	Section	Comment	Response
Josephine Grima	Evidence Summary	The Marfan Foundation represents a group of genetic conditions of connective tissue (Marfan syndrome, Loeys-Dietz Syndrome, and Vascular Ehlers Danlos Syndrome, that have life-threatening vascular events which often result in loss of life in the emergency room because of variability in diagnostic performance for these rare conditions. In most cases, this results from inadequate knowledge for the need for timely imaging to determine life-threatening vascular events especially in young or middle-aged individuals. Therefore, we strongly agree that there is enormous opportunity for diagnostic improvement of these rare conditions especially when a known genetic predisposition exists. We would applaud system wide scalable solutions to combat the misdiagnosed clinical presentations and provide enhanced diagnostic expertise at the bedside, solution sets needed to capitalize on training, teamwork and technology. We would hope that these training modules contain key components of genetic predisposition for these conditions that could result in 6 of the 10 most misdiagnosed related harms (stroke, myocardial infarction, aortic aneurysm and dissection, venous thromboembolism, sepsis, and arterial thromboembolism. In a time when there is an emphasis on personalized medicine due to specific genetic make-up, the ER field should be more up-to-date on recognizing the need for urgency of care for these conditions.	Based on the results of our systematic review, we concur that life-threatening vascular events (including those associated with genetic conditions such as Marfan syndrome and other inherited connective tissue disorders) should be an important area of focus for ED quality improvement.

Commentator & Affiliation	Section	Comment	Response
Josephine Grima	Introduction	We feel that life-threatening genetic aortic and vascular conditions should be one of the major focus areas for new quality improvement programs to all those decision-makers at every level (individual clinicians, ED directors, hospital safety officers, national policy makers, etc.) in order to reveal critical insights about diagnostic failures in the hopes to improve ER outcomes.	Based on the results of our systematic review, we concur that life-threatening vascular events (including those associated with genetic conditions such as Marfan syndrome and other inherited connective tissue disorders) should be an important area of focus for ED quality improvement.
Josephine Grima	Methods	While it is hard to comment on the methodology used in this paper, it should be noted that the study used patient population contributors to examine error such as age, sex, race, and ethnicity. However, known genetic predisposition was not taken into consideration and can be considered a weakness of the current study. This could be an area of knowledge, which would make significant improvement to outcomes in the ER, by reducing response times to imaging to rule out life-threatening vascular events. This could be easily determined through patient records, genetic testing, and family history to ensure timely and appropriate intervention, which could save a life.	To our knowledge, genetic risk factors (as direct "patient characteristics" that predict the likelihood of diagnostic error) or the failure to obtain a clinical genetic/family history (as a clinical process variable leading to misdiagnosis) were not specifically called out and measured in any of the included studies that addressed overall diagnostic error.

Commentator & Affiliation	Section	Comment	Response
Josephine Grima	Results	Our organization can attest to the results found in the study identifying 10 major causes of misdiagnosis since the genetic aortic and vascular conditions The Marfan Foundation represents often results in misdiagnosis in the ED resulting in loss of life. We believe that Marfan, Loeys-Dietz, Vascular Ehlers-Danlos syndromes present in highly variable ways from mild to severe but can encompass as many as 6 of the top 10 issues which cause harm and are prime areas for misdiagnosis (stroke, myocardial infarction, aortic aneurysm and dissection, venous thromboembolism, sepsis, and arterial thromboembolism). In an age where there is an emphasis on personalized medicine due to specific genetic make-up, the ER field should be more up-to-date on recognizing the need for urgency of care for these conditions. Additionally because of the rarity of these conditions, robust data on the mishaps in the ER can be lacking.	We were able to identify a few studies that called out a history of Marfan syndrome as a factor in attempting to predict aortic dissection misdiagnosis. We have added text to the "patient characteristics" section of KQ3 ("None of the studies on risk of delays in aortic dissection diagnosis found a statistically significant difference between those with a history of Marfan's syndrome and those without, although the presence of a known history was, if anything, protective (median time from presentation to diagnosis 2.2 hours for those with a known history versus 4.5 hours for those without, p=0.066).")

Commentator & Affiliation	Section	Comment	Response
Josephine Grima	Conclusion	The Marfan Foundation hopes that our comments can shed some light on the outcomes in the ER for genetically predisposed individuals for vascular events such as life-threatening aneurysms and strokes. We are interested in working with organizations, agencies, government, and other entities that can help raise awareness of this group of disorders. We agree with the conclusions that system-wide, scalable solutions need to be developed to tackle cognitive problems, and that these solution sets must be targeted to address the most commonly misdiagnosed clinical presentations. Genetic aortic and vascular conditions are key conditions to provide opportunities for learning and urgency of care in the ER. We would be in support of reliable delivery of enhanced diagnostic expertise at the bedside, solution sets which utilize training, teamwork, and technology. We are willing to participate in any way possible to reduce poor outcomes in the ED. We have the best medical advisors in the field and are hoping to make a dramatic impact on ER knowledge so that the lives of so many in our member communities can be saved. The proper ER evaluation is key for timely intervention that can save lives.	We thank the Marfan Foundation for their interest in the topic and advocacy on behalf of improved ED diagnosis.
Anonymous	Introduction	Might be interesting to compare the costs of building systems and the associated evaluations to see what the offset would be from the estimated current costs to the US healthcare system.	This is an interesting concept but falls outside the scope of the systematic review.

Commentator & Affiliation	Section	Comment	Response
Stephen Raab	General	This draft effectiveness review should be lauded for its extensive and ambitious evaluation of the field diagnostic error in emergency medicine.	Thank you for the kind words.
Stephen Raab	General	The authors highlighted the current state of diagnostic error research as well as identified knowledge and practice gaps that would foster additional studies. As the authors suggest, the lack of definitions has created uncertainty in understanding the diagnostic process, diagnoses, and errors in diagnosis. The National Academy of Medicine (NAM) defined diagnostic error as the failure to provide an accurate and timely explanation of the patient's health problem(s) or the failure to communicate that explanation to the patient. This definition is unclear and incomplete. The approach to study diagnostic error is not standardized, which leads to different definitions and study methods.	We agree that the NAM definition is incomplete in the sense that it does not describe in detail how these parameters should be measured, which, in turn, leaves room for variation in measurement. For this reason, one of our central policy considerations is the standardization of measurement for the purposes of research reporting as well as operational benchmarking and public quality reporting.
Stephen Raab	General	It would be helpful if the researchers had first defined emergency medicine practice in terms of diagnostics with referrals, admissions, and discharges. In what scenarios were diagnoses actually made?	Most of the literature addressed discharged patients; several papers addressed admissions; and only one paper directly addressed both admitted and discharged patients (Calder, 2010).

Commentator & Affiliation	Section	Comment	Response
Stephen Raab	General	This systematic review did not address all emergency medicine visits, although the literature search focused on reporting diagnostic errors or mis-diagnosis related harms. The reported results largely described the causes of serious mis-diagnosis related harms. This frame is relatively narrow and does not directly address the commonly used frame of overall diagnostic accuracy or precision. As a result, it is difficult to compare these error data to other fields in medicine.	The report included studies of all emergency medicine visits (these are reported in responses to KQ1a-b-c, KQ2a-c, and KQ3a). However, we supplemented this analysis with additional disease-specific literature for the diseases associated with the most commonly harmful diseases. The report does address the overall diagnostic accuracy frame (KQ2a, KQ2c). The data are quite comparable to those found in other fields in medicine, such as primary care or inpatient hospitalization (e.g., diagnostic error with or without harm – 5.7% ED, 6.3% primary care; diagnostic adverse events (any harm) – 2.0% ED, 0.7% inpatient; serious misdiagnosis-related harms – 0.3% ED, 0.4% hospital, 0.03% primary care).
Stephen Raab	General	From the frame of diagnostic testing, a medical diagnosis is a judgement or an interpretation of a disease or condition in a patient. There are two fundamental reasons for error. The first reason is a systematic tendency for a diagnosis to deviate from the true value or reference standard, which is called bias. The absence of bias is accuracy. The second reason is the propensity for a diagnosis to show scattered deviation from the true value, which is called random error. The absence of random error is precision. Diagnostic accuracy is the closeness of the diagnosis to the reference standard or “truth” and diagnostic precision is the agreement or repeatability of the test.	We agree, but clinical studies of diagnostic accuracy involving patient care (rather than radiographs, pathology slides, or other images) almost never address precision in practice, because it is generally difficult to complete inter-observer assessments of diagnostic accuracy (you cannot readily do a “second read” of the exact same patient encounter by another ED clinician). From a measurement perspective in assessing the presence or absence of error (i.e., inaccuracy), precision can be measured when the judgment of whether an error occurred is in some way subjective (e.g., chart review abstractors or raters). We have added text to the report incorporating the issue raised by this comment and addressing this point (below).

Commentator & Affiliation	Section	Comment	Response
Stephen Raab (cont'd)	General (cont'd)	(comment above)	"From a diagnostic testing perspective, a medical diagnosis is a judgement or an interpretation of a disease or condition in a patient. There are two fundamental reasons for error. The first reason is a systematic tendency for a diagnosis to deviate from the true value or reference standard, which is called bias. The absence of bias is accuracy (sometimes called "validity"). The second reason is the propensity for a diagnosis to show scattered deviation from the true value, which is called random error. The absence of random error is precision (sometimes called "reliability"). Diagnostic accuracy is the closeness of the diagnosis to the reference standard or "truth" and diagnostic precision is the inter-observer agreement or repeatability of the test (in this case, a clinical diagnosis). Studies of diagnostic error in radiology, pathology, or other image-based fields are readily able to assess clinical precision because the specific clinical artifact that is the subject of diagnosis (radiograph, histopathology slide, etc.) can be re-examined by a second clinician without loss of fidelity. However, studies of diagnostic error in the ED (or any other clinical practice setting involving a typical, multi-faceted patient encounter) rarely, if ever, can do so—the full clinical counter (as experienced by the first clinician) cannot readily be reproduced. Thus, the ED-based studies assessed in KQ2 (error rates) nominally focus on diagnostic accuracy (relative to some reference standard [presumed] "true" diagnosis), not diagnostic precision (relative to a second "equivalent" observer). Accordingly, precision (in the inter-rater or test-retest reliability sense) plays only a minor role in this report and only at a "meta" level—specifically, some research studies report the measurement precision of assessing clinical accuracy (e.g., if chart review was performed by two independent human raters judging the presence or absence of a diagnostic error, misdiagnosis-related harm, or preventable harm). However, it should also be noted that judgments of diagnostic error (often called "inaccuracy") usually do not help us distinguish between systematic (bias) and unsystematic (random) error."

Commentator & Affiliation	Section	Comment	Response
Stephen Raab	General	The three main sources for variation in a diagnostic test are the patient, the testing process, and the observer (i.e., diagnostician). Variation contributes to bias and random error. Diagnosticians not only are the observers who make diagnoses but also are part of the testing process (e.g., by obtaining clinical history or performing a physical examination).	We agree, and these issues are addressed in KQ3 (1. patient demographic/illness characteristics, 2. facility or context-specific systems factors, and 3. clinician characteristics). We have added text to the report incorporating the comment and addressing this point (below). "The three main sources for variation in a diagnostic "test" (in this case a clinical diagnosis rendered by the ED care process) are the patient, the testing process, and the observer (i.e., diagnostician). Variation contributes to bias and random error. Diagnosticians are not only the observers who make diagnoses but also are part of the testing process (e.g., by obtaining clinical history or performing a physical examination). As part of our study method for this report, we prospectively defined characteristics and factors that have been shown to impact diagnostic errors in prior studies (Table 13) and used these to abstract data from included studies. Individual clinicians were rarely the subject of research on diagnostic error, so variation at the level of clinicians reflects "average" characteristics among a pool of clinicians within a given study."
Stephen Raab	General	In clinical practice, disease reference standards are insufficiently understood, developed, and implemented, and diagnosticians disagree on patient diagnoses.	We agree and have added text to the report incorporating the issue from this comment into the Limitations.

Commentator & Affiliation	Section	Comment	Response
Stephen Raab	General	Safety-I science involves the detection and prevention of error. A common method of error detection is blinded secondary case review, which shows diagnostician agreement or disagreement in the diagnosis. In order to assess a base line error frequency, the case sample must be representative of the patient population. Obtaining a representative sample allows for a detailed study of error and safe practice attributes. The study of safe practices is Safety-II science, or resilient practice. The failure to study baseline safety and error frequency lowers the understanding of variance, which in turn, limits the ability to design error prevention or safe strategies. In Safety-I, diagnostic accuracy is expressed as a percentage and precision is often expressed as a kappa statistic, reflecting the level of agreement.	Clinical studies of diagnostic accuracy involving patient care (rather than radiographs, pathology slides, or other images) almost never address precision in practice, because it is generally difficult to complete inter-observer assessments of diagnostic accuracy (you can't readily do a "second read" of the exact same patient encounter by another ED clinician). From a measurement perspective in assessing the presence or absence of error (i.e., inaccuracy), precision can be measured when the judgment of whether an error occurred is in some way subjective (e.g., chart review abstractors or raters). We have added text to the report incorporating the issue raised by this comment and addressing this point (below).

Commentator & Affiliation	Section	Comment	Response
Stephen Raab (cont'd)	General (cont'd)	(comment above)	"From a diagnostic testing perspective, a medical diagnosis is a judgement or an interpretation of a disease or condition in a patient. There are two fundamental reasons for error. The first reason is a systematic tendency for a diagnosis to deviate from the true value or reference standard, which is called bias. The absence of bias is accuracy (sometimes called "validity"). The second reason is the propensity for a diagnosis to show scattered deviation from the true value, which is called random error. The absence of random error is precision (sometimes called "reliability"). Diagnostic accuracy is the closeness of the diagnosis to the reference standard or "truth" and diagnostic precision is the inter-observer agreement or repeatability of the test (in this case, a clinical diagnosis). Studies of diagnostic error in radiology, pathology, or other image-based fields are readily able to assess clinical precision because the specific clinical artifact that is the subject of diagnosis (radiograph, histopathology slide, etc.) can be re-examined by a second clinician without loss of fidelity. However, studies of diagnostic error in the ED (or any other clinical practice setting involving a typical, multi-faceted patient encounter) rarely, if ever, can do so—the full clinical counter (as experienced by the first clinician) cannot readily be reproduced. Thus, the ED-based studies assessed in KQ2 (error rates) nominally focus on diagnostic accuracy (relative to some reference standard [presumed] "true" diagnosis), not diagnostic precision (relative to a second "equivalent" observer). Accordingly, precision (in the inter-rater or test-retest reliability sense) plays only a minor role in this report and only at a "meta" level—specifically, some research studies report the measurement precision of assessing clinical accuracy (e.g., if chart review was performed by two independent human raters judging the presence or absence of a diagnostic error, misdiagnosis-related harm, or preventable harm). However, it should also be noted that judgments of diagnostic error (often called "inaccuracy") usually do not help us distinguish between systematic (bias) and unsystematic (random) error."

Commentator & Affiliation	Section	Comment	Response
Stephen Raab	Abstract	In the Structured Abstract, the Objectives were to perform a "systematic review to determine the most frequent diseases and clinical presentations associated with diagnostic errors and resulting harms in the emergency department (ED) and the factors associated with these errors." This frame did not assess a general diagnostic error frequency and consequently the level of diagnostic accuracy and precision are not directly assessed. The frame of this report is on finding the most common presentations, diseases, and harms associated with diagnostic error. Although the findings reported are significant, the ability to understand variance, frequency, causation, and prevention are limited. The inclusion of malpractice claims provides insight into a specific category of error with potential causes and harms but does not provide a baseline of practice outside of the medical-legal setting. The inclusion of trigger detected errors is interesting, but biased on inclusion and exclusion variances.	We agree that first sentence of the abstract was incomplete and have reworded per below to make clearer that we did, in fact, assess "general diagnostic error frequency." Only KQ1a and KQ3a relied heavily on numerator-only studies. Malpractice claims (and other numerator only data) were excluded from consideration for KQ2 on error and harm rates. To clarify data sources, we have added a new Appendix (Table A-1) to the report that identifies data types and sources by Key Question. Revised text in Abstract: "Objectives. We conducted a systematic review to determine the most frequent diseases and clinical presentations associated with diagnostic errors (and resulting harms) in the emergency department (ED), measure error frequency, and assess causal factors associated with these errors."
Stephen Raab	General	The medical literature contains hundreds of interobserver variability studies in emergency medicine activities, pathways, and hand-offs. These studies were not considered, although imprecision inherently is embedded in the errors collected in this review and important for quality improvement activities.	A specific example or two of a study/studies that examined clinical diagnostic errors in an "interobserver variability" framework would have been helpful to better respond to this critique. We did not restrict our search to studies of "diagnostic accuracy" (or exclude "precision" or "variability" studies). Instead, we sought to capture all potentially relevant studies that addressed diagnostic error in the emergency department. Thus, any studies meeting these criteria were captured by our search strategy and would have been included if they reported on diagnostic error. A search for variation in every individual ED process (e.g., handoffs) was beyond the scope of the review.

Commentator & Affiliation	Section	Comment	Response
Stephen Raab	Evidence Summary	The evidence summary reports that half of all serious harms occur in 10 diseases. As prevalence of disease is not assessed, the frequency of serious harms is not considered.	The frequency of serious harms (i.e., NAIC 6-9 --- death or permanent disability at the level of loss of one limb or worse) is addressed in KQ2a and is measured at about 0.3% of all ED visits.
Stephen Raab	Results	The authors indicated that "root causes of ED diagnostic errors were disproportionately cognitive in nature and mainly happen at the bedside." The RCA methods were not discussed. In some diagnostic error scenarios, variance in process steps is augmentative based on contributions from multiple sources (observer, patient, and testing process). The KQ 3 section raises the challenges of how to assess variance and error cause in certain diseases. It is not surprising that emergency medicine diagnostic errors have a cognitive component, as these errors are often detected through secondary review methods that show imprecision in cognitive diagnoses.	In the text of KQ3, the method of root cause analysis was described briefly as follows: "According to the published study, "relevant factors in each case are abstracted based on a complete review of the medical and legal case file including case summaries, medical record data, depositions, and legal proceedings. Cases are reviewed and coded by experienced clinical taxonomy specialists (typically registered nurses with at least 10 years of quality or risk management experience), who abstract data using a multi-tiered coding taxonomy."" As noted in Figure 16, the mean number of cause categories per case was 2.4, so we concur that often more than one cause is present.

Commentator & Affiliation	Section	Comment	Response
Stephen Raab	Results	In the Results section, the authors estimated a weighted average overall diagnostic error rate of 5.6% per ED visit by combining the error rate among ED discharges (4.1%) from a case-control study at a large university hospital in Spain with the error rate among ED admissions (12.3%) from a rigorous, prospective study at a university hospital in Switzerland. Combining two single institutional studies is far from ideal and not generalizable. The authors use this frequency to estimate US error rates. For example, the researchers used the estimated average diagnostic error rate across all diseases (5.6%) as a comparative measure to myocardial infarction false negative rates of 1-2% and top harm-producing dangerous disease false negative rates of 14-28%. These estimated rates are not comparative. It might be more appropriate to perform a sensitivity analysis until additional studies are completed that directly measure this frequency.	Obviously meta-analyses are, by their nature, imperfect (as with all scientific studies). By definition, rates (or other forms of data) are aggregated and mathematically synthesized across data sources from different studies on the same topic that are necessarily non-identical. This is both a weakness and a strength. It is a weakness because comparability is not ensured. It is a strength because variation in study designs across centers is "averaged" out across studies. We limited quantitative synthesis to sufficiently similar studies (i.e., we do not combine studies of different designs that draw from different source populations). We were also careful to explain why the rates analyzed and summarized are internally consistent and coherent ("…the 4.1 percent estimate for the ED diagnostic error rate is correctly positioned within the spectrum of error/harm frequencies—diagnostic errors among admitted "non-specific" symptom cases (54%) > diagnostic errors among admitted patients (12%) > diagnostic errors among treat-and-release discharges (4%) > diagnostic errors resulting in adverse events (2%) > diagnostic errors resulting in serious harms, including death or permanent disability (0.3%). Finally, the overall error rate of 5.7% is comparable to that found in rigorous US-based studies of other frontline care settings (e.g., 6.3% overall diagnostic error rate in US-based primary care clinics).17"). Nevertheless, we have added to the Summary and Conclusions a more direct statement of Limitations in the underlying data sources ("Our review findings are tempered by limitations in the underlying evidence base, including issues linked to data sources, measurement methods, and causal relationships.").

Commentator & Affiliation	Section	Comment	Response
Stephen Raab	Abstract	In the Conclusion section of the Standard Abstract, the researchers propose developing a National Diagnostic Performance Dashboard to track performance. The researchers indicate that most errors have a cognitive component and are identified by a secondary review process, reflecting a problem in precision. Consequently, some form of adjudication would be needed to assess the level of disagreement, which may be straight forward in many cases but difficult in other cases. Currently, an infrastructure of adjudication is not existent, and diagnosticians would be challenged by this activity. Alternatively, the study and implementation of error prevention methods may facilitate diagnostician buy-in.	We do not propose that diagnostic errors be adjudicated by clinicians --- we agree this would be expensive and unwieldy. Diagnostic errors, as defined by the National Academy of Medicine, do not require demonstration of a cognitive error or process failure in diagnosis, only that the diagnosis label rendered is incorrect. Instead, we propose that the Diagnostic Performance Dashboard be used to benchmark diagnostic outcomes analogous to the Dartmouth Atlas project. As described in the report, this can be accomplished with statistically valid approaches to large data set analysis (e.g., SPADE) that require no subjective human interpretations or adjudications and are inexpensive to apply. At least for dangerous diseases requiring hospitalization, disease definitions can generally be agreed upon using a reasonable (even if imperfect) reference standard --- for example, a final inpatient hospital diagnosis of stroke is a reasonable proxy for a "true" final stroke diagnosis (see PMID: 12364739). We have now clarified in the Discussion section that we are referring to outcomes ("AHRQ, … or non-governmental organizations could monitor the overall epidemiology and variability of diagnostic performance (specifically, diagnostic outcomes, which can be adjusted for case mix severity) across the nation (analogous to the Dartmouth Atlas Project for utilization of healthcare services).")

Commentator & Affiliation	Section	Comment	Response
Pat Croskerry	General	Firstly, I believe the report does a very good job of focusing on cognition as a source of many diagnostic failures, and especially that which occurs at the bedside. However, a major concern is that it stops short of saying much about cognition. If one were asked what 'cognition' means the likely response would be that it has something to do with purposeful thinking. But there are many other features of cognition that are relevant, all of which may influence the overall complex process of thinking: critical thinking, organisation of thought, logical soundness, influence of judgment and decision making (JDM) biases. and others. Here are some observations: • Expertise in cognition mostly comes from the cognitive sciences, especially cognitive psychology. It does not come from Medicine, therefore, we should engage the input of cognitive scientists. • Generally, those who have conducted the AHRQ review will not have been trained in the cognitive aspects of clinical decision making. • Generally, the studies that have been included in this 20 year window will have been done by researchers who have not been trained in cognitive science. • How questions get asked determines what responses will be made - both by those conducting the review and by those who completed the studies.	Thank you for these thoughtful comments. We found no studies that met our inclusion criteria and directly addressed these underlying issues related to the "root causes of the root causes." We have added text to this effect to KQ3a ("We identified no studies that attempted to drill down further into the cognitive psychology of cognitive error (e.g., types of decision-making heuristics or associated cognitive biases at play)..").

Commentator & Affiliation	Section	Comment	Response
Pat Croskerry (cont'd)	General (cont'd)	• Because of the way in which medicine has been taught over the last century or so in the West, the typical response to diagnostic failure has been that the physician does not know enough i.e. their knowledge about the disease in question is incomplete. This argument has been made and continues to be made in discussions around diagnostic failure. In the present report, similar inferences are made: "Most often these (errors in diagnostic assessment) were attributed to inadequate clinical knowledge, skills, or reasoning, particularly in "atypical" cases". It is a legitimate comment in that decision making cannot be effective unless there is a sufficient knowledge base about disease, and it is clear that in some instances diagnostic failure may occur due to inadequate knowledge. • There is indeed a prevailing knowledge deficit underlying many cases of diagnostic failure, however it is not one that is being attended to. The deficit lies in a lack of awareness and understanding of how cognition works, and how it fails. Ironically, this is not usually the knowledge deficit that most people think of. It is a thinking failure about thinking failure.	(response above)

Commentator & Affiliation	Section	Comment	Response
Pat Croskerry (cont'd)	General (cont'd)	• To understand why diagnostic failure is more likely in atypical cases, we need to know something about the processes involved in human perception (pattern recognition, signal-to-noise ratio, psychophysics, manifestation of disease, fatigue, interruptions and distractions of attention etc) which are topics discussed in cognitive science, but not usually medicine. • Every discipline and most sub-disciplines in medicine have now published, in their respective literatures, clinical cases of diagnostic failure. These are not experimental - they are typically descriptive and narrative and, consequently, will not have been picked up by the search strategies used for the present review. There are about 50 but they were not cited here. Consequently, even though many physicians consider cognitive bias important, it barely gets a mention in this review.	(response above)

Commentator & Affiliation	Section	Comment	Response
Pat Croskerry (cont'd)	General (cont'd)	• Many of the studies reviewed in this study, in looking at causes of failure, have used proximal rather than distal explanations. Proximal are those closest to the error, and reflected in the DEER classification. e.g. a failure to elicit a particular piece of history, or failure to do a particular test. Those are obvious, immediate, proximal explanations. But the reasons for not taking a history or doing a test are probably due to distal explanations such as the clinician anchoring onto another explanation, or premature closure on another possibility occurred - these are cognitive phenomena (JDM biases)that explain the behaviour more distally. Proximal explanations are often obvious, visible, and measurable, whereas distal explanations are invisible, difficult to measure and have to be inferred.So, the increased difficulty that people have in generating distal explanations will tend to put them off looking for them. However, without an awareness and understanding of how cognitive biases work (often the distal explanation), most investigators will settle for the easier proximal explanation.	(response above)

Commentator & Affiliation	Section	Comment	Response
Pat Croskerry (cont'd)	General (cont'd)	• The value of doing such a review as this is that it puts the reader in a position of asking what needs to be done to tackle some of the issues raised, and what policy changes might we expect? Unfortunately, most will not be steered towards the notion that basic medical training needs to embrace cognitive science, despite several calls in the literature for this. The status quo will instead be maintained. Many medical educators themselves have not been trained in cognitive science and therefore will not see a need to provide training in it. • Concluding that cognition is involved in diagnostic failure is probably no more effective than saying fire is a major source of human injury and mortality. It is undoubtedly true but we need to know how the fire started in the first place.	(response above)