Comparative Effectiveness Review Disposition of Comments Report

Title: *Impact of Healthcare Algorithms on Racial and Ethnic Disparities in Health and Healthcare*

Draft report available for public comment from February 9, 2023, to March 9, 2023.

Citation: Tipton K, Leas BF, Flores E, Jepson C, Aysola J, Cohen J, Harhay M, Schmidt H, Weissman G, Treadwell J, Mull NK, Siddique SM. Impact of Healthcare Algorithms on Racial and Ethnic Disparities in Health and Healthcare. Comparative Effectiveness Review No. 268. (Prepared by the ECRI-Penn Medicine Evidence-based Practice Center under Contract No. 75Q80120D00002.) AHRQ Publication No. 24-EHC004. Rockville, MD: Agency for Healthcare Research and Quality; December 2023.
DOI: https://doi.org/10.23970/AHRQEPCCER268. 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. Five reviews were received from two independent reviewers, two members of the Technical Expert Panel, and one Key Informant. Several themes emerged from the reviews.

- A substantial number of comments focused on the Abstract and Executive Summary. Many of these comments raised issues about the overall approach of the review, the methodology, and the findings and are reflective of or redundant with similar comments made in the other sections of the review. We describe their concerns below, where we address those other sections (Introduction, Methods, Results) directly. But some of the comments revealed the need to improve the clarity and readability of the abstract and summary, and so we made extensive revisions to these sections to more clearly convey the key messages of the review.
- Some comments on the Introduction challenged our framing of core conceptual issues around racial and ethnic bias in algorithms, or our description of specific details of these topics. We have made significant revisions to the Introduction to address some of the concerns that were raised. We will also make additional edits prior to the public comment period to improve our description of the recommendations on estimated glomerular filtration rate (eGFR) testing made by the American Society of Nephrology (ASN) Task Force, as we recognize that we may have mischaracterized their conclusions regarding cystatin C.
- We do not intend to alter our characterization of race and ethnicity as socially constructed, and we do not think it is useful to devote a significant amount of text to discussing in depth the complex interactions between race, ethnicity, ancestry, genetics, racism, social determinants of health, and disparities. We have revised some phrasing throughout the report where more clarity might be helpful. We will also update references throughout the Introduction to ensure we capture current research and perspectives during the public comment period.
- Feedback on the review's methodology fell broadly into two categories. Some comments suggested or revealed the need for clarification of specific methodological details, and we revised the text in several places for clarity. Other comments questioned the appropriateness of our inclusion/exclusion criteria and other key aspects of our methodology, asserting that the review fails to include a wide variety of relevant literature because of the limitations we set on study eligibility. In general, our responses to such comments emphasized that our study criteria were determined after close consultation with AHRQ and our subject matter experts, and with input from the Technical Expert Panel (TEP) and Key Informants (KIs). We also cited the infeasibility of reviewing a much larger body of literature within a constrained timeframe.
- Input on the Results focused in large part on our description of the methods and findings of specific studies. We have substantially revised the Results and determined that several studies that had been previously evaluated for both Key Questions (KQs) were better suited to KQ 2 only. We streamlined the narrative description of the evidence in KQ 1 and attempted to clarify specific points raised by the peer reviewers about individual studies.

- To make the large volume of information in the Results more digestible, we replaced the previous Tables 3 and 4 with a revised and expanded Table 3. We also inserted brief descriptive tables at the outset of each clinical area addressed for KQ 1 to better introduce those sections.
- We received some feedback on the Contextual Questions (CQs) as well. In response to a few comments about the evidence supporting our findings for CQ 1–3, we added text to highlight conclusions that were based primarily on our discussions with the TEP and KIs. We also revised the results described in CQ 4 to provide better clarity.
- Some comments pointed to the need for the Discussion to be clearer, more consistent with the Results, and more actionable. We made some revisions to the Discussion and added a section of specific recommendations for various stakeholder groups. We will continue to refine the Discussion during the public comment period.
- Finally, we received feedback indicating that the report did not adequately address current research on artificial intelligence (AI) tools. We explain in our responses below that the scope of this project focused on algorithms and mitigation strategies related to clinical care. Our subject matter experts in clinical prediction modeling advised us that some of the approaches used in the broader field of AI to detect possible bias are not, at least as of now, commonly used with clinical algorithms. For CQ 2, we did look more generally at possible mitigation strategies, but we curated the approaches presented to ensure maximal relevance for clinical medicine. Moreover, we focused mainly on strategies that can be implemented to mitigate an existing bias rather than automated tools that are designed primarily to detect possible bias.

Peer, Key Informant, and Technical Expert Panel Comments and Author Response

Commenter	Section	Comment	Response
TEP/KI Reviewer #1	General	This is a very comprehensive review that appropriately illuminates a number of nuances in the development and implementation of algorithms that use race. As former director of one of the original AHRQ EPCs 25 years ago, I congratulate and commend the team on their work.	Thank you for your review.
TEP/KI Reviewer #1	General	I liked the expansion of the topic of other algorithms that do not include race and may incite health disparities. An important point that should be emphasized is the algorithms are not likely to be the primary drive of health inequities and energy to eliminate them might be more effectively put on other important drivers.	Thank you for this suggestion. We will consider how to incorporate this point into the Discussion, during the public comment period.
TEP/KI Reviewer #1	General	Many of the algorithm changes that are done change the reporting of the algorithm rather than redo the calculation and when they do this use WHITE as the reference standard as if the Black persons data was tainted. This is a trap that many advocates do not realize. We did this in genomics for a while where initial studies identifying alleles in White Europeans were considered as normal and alleles in Africans or other populations were considered “mutant” alleles. I wish the report would emphasize this about the changes to the algorithms that were made and therefore tested and made it into this review.	We are carefully examining how to include this context, and will make additional edits during the public comment period.
TEP/KI Reviewer #1	General	I know the most about the kidney area and my comments are very important for the evidence and how to view it in this arena. Again, congratulations on this wonderful piece. It will be very important to this debate.	Thank you.
Peer Reviewer #2	General	The report is clinically meaningful, but there is a fundamental issue with the approach. Current AHRQ procedures are usually focused on requiring a high level of evidence to establish the benefit of an intervention. In this case, AHRQ is applying this same high level of evidence to establish the harm of interventions already in wide use. This does not comport with the precautionary principle applied in fields such as environmental health, where lesser levels of evidence are acceptable to generate warnings about safety while further data is gathered. Requiring a prospective cohort or randomized control trial to establish harm allows harm to continue being perpetuated while waiting for resource intensive research that may never materialize.	The requested scope for this review was to evaluate the use of algorithms in creating disparities. We have now modified our report to explicitly state the implications as well as potential limitations of taking this approach.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	General	The literature review is extremely narrow and inadequate. Assumptions and conclusions are made that do not merit what is known in the field and contradicts the findings sited. The key questions and context questions were not adequately answered given what is known in the field and in the literature. ES-4 line 39 summarizes the problem with the review. It is not generalizable beyond RCT, which is not typical of the algorithm bias literature because it is not used. References are outdated, especially those in the background section regarding race and ethnicity. This section should be totally revised with more balanced perspectives and current information. The selection of articles has the same fatal flaw of disparate number of other racial/ethnic groups compared to the 80-90% of whites in the data. This creates a challenge for algorithms to function properly. This is such an issue that AI developed synthetic data. Yet, these authors repeatedly assess the articles without identifying the disparate N for racial/ethnic groups as a reason for the algorithms challenge to perform.	We appreciate your careful review and thoughtful comments. However, there are certain statements in this comment that leaves the research team unclear as to how to respond. For example, the first part is in reference to the entire literature and does not directly clarify what assumptions and conclusions lack merit. The following comment about generalizability beyond RCTs does not provide actionable guidance as prediction models are almost always in cohort studies, and RCTs are not the focus here. Finally, we recognize the important impact of disparate populations in algorithmic training data, and address this in in report.
Peer Reviewer #1	General	There are no conclusions drawn for each question based on the analyses provided in the section that matches the literature. It is difficult to determine answers. Although the limitations are noted, stressing the fact that these results should not be the foundation of decisions because of the limitations are not stressed enough.	Thank you for your feedback. We have substantially revised the Results and Discussion.
Peer Reviewer #1	General	BIPOC should not be used ever!!!	Our Subject Matter Experts, with expertise in health equity, disparities, and minority health, advised us that BIPOC is an appropriate term for this report. While we appreciate that different journals, organizations, and professional societies maintain different recommendations, we are not aware of one that suggests eliminating the use of the term BIPOC. Indeed, it is commonly used in journals that publish relevant content.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	General	Font should all be the same. References should follow the same format.	We have followed the style and formatting standards that are required for this report as codified in AHRQ's Publication Guide.
TEP/KI Reviewer #2	General	The methods were well described and sound. The general conclusions of the report are not a surprise. As there is a general shortage of literature in this space- I would have liked more analysis on some of the articles identified but not included, for example, studies from other countries and studies that did not include outcomes could both have been analyzed and summarized for findings.	While we agree that such articles could provide useful insights, the scope and timeline of this project made it infeasible to review and summarize such studies, even in a very general manner. Additionally, through our team's professional interaction with international colleagues we have found that the focus of algorithmic fairness specific to race seems to be predominantly centered in the US.
Peer Reviewer #1	Abstract	Vi and vii - In the entire Results section there is no mention of data as a reason for algorithm biases. This is a known fact….not enough images, missing data, etc	While we agree that data sources are a key driver of algorithmic bias, data sources are infrequently studied in this context when clinical prediction models are trained and validated. Thus, they are not reported as identified reasons in the papers presenting clinical prediction models. Rather, we identified several data-related mitigation strategies (e.g., input variables, output variables, population represented in the training data) that we do report in the Results section of the structured abstract.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Abstract	Vi line 15 It is best to use third person when writing science. Remove we and use "Published and grey literature….2022 was searched / Vi line 18 remove we and restate in third person / Vi line 22 remove we and restate in third person etc Throughout the entire report – Please remove personal pronouns and state in third person. Science reports are written in third person….it is not personal.	We acknowledge that use of the passive voice is a long-held convention in scientific writing. However, avoidance of the first person voice is no longer universal; for example, APA style now advocates use of the first person. Our EPC reports routinely employ first person voice when it aids clarity, brevity, and directness.
Peer Reviewer #1	Abstract	Vi line 48 - there is lack of evidence because the articles address development, not implementation of the AI tool. To find evidence, the literature has to focus on the implementation results.	We agree that results of implementation studies are ideal, but our systematic search identified very few such studies.
Peer Reviewer #1	Abstract	Vi line 50 - There is no documentation for the sentence that begins with Evidence suggests…..this is not true and is one of the reasons that AI leads to disparities – see the study of AI use for payments by Kaiser.	That severity of illness scores can overestimate risk among Black patients is well established in several studies, including some that rely on data from the Kaiser health system (See among others, Ashana et al). Without a specific citation from the Reviewer, it is impossible to provide a specific response to the content of the mentioned study.
Peer Reviewer #1	Abstract	Vii – conclusion makes assumptive statements without evidence ….it needs to reworked entirely.	The abstract has been revised substantially.
Peer Reviewer #1	Executive Summary	ES-4 lines 29-38 It might be better to restate these results as "The method of review and assessment used discerned no…." instead of "we discerned" – the entire paragraph should reflect that tone instead of a tone that nothing was found. These results are a result of the way the analyses were conducted---not what is.	The sentence has been removed, and the paragraph extensively rewritten.
Peer Reviewer #1	Executive Summary	ES4- line 30 there should be a comma after indicating / before which	That sentence was removed.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Executive Summary	ES4-line 39 this sentence is highly significant for the generalizability of these findings and how the methodology was conducted. If this is the case, then there is no need for RCT focus of assessment.	Thank you for your feedback. One of our main conclusions is that there are very few studies, RCT or otherwise, that adequately assess the impact of algorithms on racial and ethnic disparities.
Peer Reviewer #1	Executive Summary	ES4-line 43 there are studies in the literature that do such, but were not included in this assessment	We acknowledge that many studies that address this topic were excluded, but we included all studies that met our specific criteria for this report. These criteria represent our best effort to identify appropriate studies within a feasible scope of work, and were carefully reviewed by our Subject Matter Experts, Key Informants, Technical Expert Panel, and AHRQ.
Peer Reviewer #1	Executive Summary	ES4- line lines 44-48 are referring what is known in the field of health disparities as upstream effects that lead to such. This paragraph should be reworded to reflect the field knowledge and what is occurring when algorithms have biases.	This paragraph was removed during our extensive revisions to this section.
Peer Reviewer #1	Executive Summary	ES4-line 51 DO NOT USE BIPOC— the racial/ethnic groups can be spelled out or underrepresented minority populations can be use.	Our Subject Matter Experts, with expertise in health equity, disparities, and minority health, advised us that BIPOC is appropriate for use in this report.
Peer Reviewer #1	Executive Summary	ES4- line 55 Is a recommendation not an assessment. The statement is also false. We know the outcomes and strategies to fix algorithm bias. This statement is a result of the lack of comprehensive review of the literature.	This sentence was removed.
Peer Reviewer #1	Executive Summary	ES-3 lines 5-10 have been stated numerous times and says nothing of significance of why they were include or not, nor how or why they contribute to the key questions.	The paragraph has been extensively revised.
Peer Reviewer #1	Executive Summary	ES-3 line 13 this statement needs to be highlighted “conclusions are limited”	We revised the executive summary to improve clarity.
Peer Reviewer #1	Executive Summary	ES-3 line 13-16 this is an erroneous over statement. Please explain how your conclusion was derived.	We revised the executive summary to improve clarity.
Peer Reviewer #1	Executive Summary	ES-3 line 18-20 please explain how and why this conclusion was drawn.	This paragraph has been extensively revised.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Executive Summary	ES-3 lines 23-30 should be written such that each type is associated with whatever disease/disorder was the focus. Also, ROB should be associated as well. The way the paragraph is written offers no understanding of results or the impact of algorithms.	We did not go into this level of detail in the Executive Summary, which is designed to be very concise. These issues are treated more comprehensively in the Results.
Peer Reviewer #1	Executive Summary	ES-3 lines 32-38 again these numbers mean little out of context of the algorithm and what was the bias. It is data without meaning.	We have revised this paragraph, but note also that, given the very broad scope of the review, the Executive Summary can only briefly introduce the findings.
TEP/KI Reviewer #1	Executive Summary	Page 12 , Line 38-41 Evidence suggests that removing a race coefficient from eGFR by dropping it an existing equation results in significantly more diagnoses of chronic and severe kidney disease in Black patients,which can then lead to increased eligibility for kidney transplant and alternatively in underuse or underdosing of important chemotherapy, antidiabetes and pain medications and underenrollment of Black persons in clinical trials.	This sentence was revised in accordance with this suggestion.
Peer Reviewer #1	Executive Summary	ES-3 line 38-41 this sentence cannot be assumed by the sentences above in the paragraph. This is a dangerous overstatement of mitigation strategies – Provide appropriate context or remove. This will cause more harm than good.	This paragraph was revised to add important context. We agree that removal of race in eGFR is not a universal model for mitigating bias in all cases, and does not even mitigate all risks for patients with kidney disease. We have sought to clarify this throughout the report.
Peer Reviewer #1	Executive Summary	ES-3 line 42-48 provides 3 disjointed separate comments. The first sentence is somewhat true that the mitigation strategies are often based upon algorithm testing strategies, typically limited to known strategies in python. However, no strategies or techniques were noted to make the assumption. The second sentence is somewhat true as well but clinical outcomes are not the only algorithms that need real world testing. Clinical outcomes are probably the least likely because these are often testing the results of medications, which cannot be known until consumed by large pools of participants. The last sentence is indeed true but it is out of context and provides no relation to the two sentences prior in the same paragraph.	The first sentence has been revised to refer specifically to the studies included in this report. With regard to clinical outcomes, we agree that other algorithms require real world testing as well; however, this review was designed, with input from AHRQ, to focus on algorithms that are associated with clinical outcomes.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Executive Summary	ES-3 line 49-51 This is a misstatement. The scope of AI can be assessed, and it does span across the spectrum of AI and entering realms of use from clinical to payer to prevention thru diagnostics to treatments. It also covering tracking and monitoring as well as reminders for health care interventions. There are no references to public perceptions of algorithms – How were these conclusions drawn and placed in another disjointed paragraph of one liners that are disconnected?	We agree that the scope can be assessed in broad and general terms, and we sought to describe that in the report. But this point is meant to convey that is it very difficult to quantify that scope, with, for example, estimates of how many clinical algorithms in current use include race, or how many patients are affected by potentially biased algorithms. As for our conclusions about public perceptions, these are based on input from our Key Informants, Technical Expert Panel, and Subject Matter Experts.
TEP/KI Reviewer #1	Executive Summary	Page 11 Some of the language starts with the premise that algorithms with race are causal of disparities, rather than a hypotheses with exquipoise and references cited are based on perspectives or viewpoints. An example with more equipoise is the following. “but because race and ethnicity are socially constructed, it has been alleged that their inclusion may exacerbate or perpetuate health and healthcare disparities due to structural biases and racism in healthcare.[4-6] There are many ways to remove race from algorithms as outlined by Powe N JAMA with different effects and there is evidence of bad effects from the practice employed by many insitutions.	We have made the suggested change to the wording of the sentence.
Peer Reviewer #1	Executive Summary	ES2 line 7 – race and ethnicity collection does not cause health disparities – this is blaming the victim for the victimization. Please remove that sentence. Whether r/e is a social construct or not, it is not the cause nor does it perpetuate health disparity and racism in health care	We revised this sentence, and maintain that one of the primary premises of this report is that inclusion of race and ethnicity as an input variable in clinical algorithms can result in perpetuation and/or exacerbation of disparities in health and healthcare outcomes.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Executive Summary	ES2 line 39 – implementation identification of biases can be determined by other means that empirical	After a preliminary review of the literature, we determined that restricting Key Question 1 to empirical studies was a necessary step to enable completion of this report within the defined timeline. Including a broader range of studies would have required far more time and resources than we had available.
Peer Reviewer #1	Executive Summary	ES2 line 42-47 – there are many ways to test for biases…what methods did they use regarding the testing and retesting of inputs and with/without race/ethnicity—using one method, such as ROB, is not sufficient.	We assume that by "they" the reviewer is referring to the authors of the included studies and is asking what methods those authors used to test for bias in the algorithms. That is not exactly what the passage in question is about; it is a list of the categories of comparators and outcomes that we considered to be of interest. It is true that, in Key Question 1, the method used to test for bias will be reflected in the type of comparator selected (in Key Question 2, it would be reflected in the type of intervention selected). Two categories of comparator listed in the passage (same algorithm with or without race, same algorithm with or without other variables that may contribute to bias) seem relevant to the concerns expressed. The meaning of the reference to ROB is unclear; it may possibly refer to the sentence immediately following the passage in question, which describes the method we used to assess ROB of the studies (not of the algorithms being studied).

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Executive Summary	ES2 line 51-54 are statements – it does not explain the methodologies that were used.	To improve the brevity of the Executive Summary, we did not describe the methodology in detail.
Peer Reviewer #1	Executive Summary	ES1 Exec Summary - Remove all personal pronouns and write in third person. For every bullet point, it would be more useful to state how the algorithm was assessed that lead to the conclusion statement that is detached and rather meaningless. Apply this to bullet 2 line 17; bullet 3 line 25	The bullet points have been extensively revised for clarity. The use of first person language is a convention we frequently use in EPC reports.
Peer Reviewer #1	Executive Summary	ES1 line 32 –what was the conclusion or a statement that pulls the finding together	This bullet point has been extensively revised.
Peer Reviewer #1	Executive Summary	ES1 line 35- makes no sense to the prior sentence. – this study main issue was allocating higher priority to people who donated a kidney, which ended up being white women.	This bullet point has been extensively revised.
Peer Reviewer #1	Executive Summary	ES1 line 40 - the examples do not make sense as contextual factors—this needs to be clarified—it's a hodge-podge of unrelated points – r/e should be used in algorithms or not reviewed for this project; representatives in clinical studies means what?—means there were what r/e groups included ---what other characteristics would matter when assessing algorithm bias and health disparities	This sentence was removed.
Peer Reviewer #1	Executive Summary	ES1 line 47 - the sentence "more primary research …..is needed." Is a recommendation not a statement in the exec summary regarding the spectrum of healthcare	This sentence was removed.
Peer Reviewer #1	Executive Summary	ES1 line 49 – This is regulatory and standards have been developed –What KQ or CQ does this comment actually involve.	This sentence was removed.
Peer Reviewer #2	Introduction	Some clarity could be gained, however, with greater consistency in terminology: race/ethnicity or race and ethnicity. Currently the paper switches back and forth. JAMA's updated guidance on the reporting of race and ethnicity may be helpful in this respect (see https://jamanetwork.com/journals/jama/fullarticle/2783090).	We have modified our terminology to consistently use "race and ethnicity".

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #2	Introduction	Also recommend the authors revise the capitalization of “White” (recognizing that there are varying guidelines on this, but arguably the AP guidelines have been most widely accepted: https://apnews.com/article/archive-race-and-ethnicity-9105661462) and the use of “disparities” instead of “inequities”.	We disagree with making White non-capital while capitalizing Black or BIPOC populations, as it reinforces the idea of White as the norm, default, or reference group. With this report we are aligning with the presentation of language that does not reinforce cis-heteronormative White individuals as the norm, from which all others deviate from. https://cssp.org/2020/03/recognizing-race-in-language-why-we-capitalize-black-and-white/ Lastly, we have included rationale for use of disparities versus inequities and definitions to orient readers and note that when summarizing studies, we are also constrained by what terms and demographic categories the study authors themselves utilize to characterize differences if insufficient context is provided.
Peer Reviewer #1	Introduction	Introduction references are outdated and should be revised to include more current studies, especially those that categorize race and ethnicity in the current OMB Directive 15 Standard. There is an imbalance or a misrepresentation of race and ethnicity health outcomes and justifications for health disparities. This section should provide a more comprehensive view of health disparities, including the information. The kidney examples lack comprehension of why changes were made to the original algorithm, which reflects the lack of understanding of the context. The way the highlights are presented in the introduction does not get at the biases in algorithms or the potential harms done.	We are reviewing all the references included in the Introduction and will add updated citations during the public comment period. We have also revised the Introduction for clarity.
TEP/KI Reviewer #1	Introduction	Thorough, well structured	Thank you.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
TEP/KI Reviewer #2	Introduction	Good- however, I found myself needing to go back many times to clarify the difference between contextual questions and key questions.	We recognize the complexity of addressing numerous questions in different sections of the report, and will continue to improve clarity and readability during the public comment period.
TEP/KI Reviewer #1	Introduction	There need to be more equipoise and nuance in the language about effects of removing race by simply dropping the race coefficient. See Diao JA, Wu GJ, Taylor HA, Tucker JK, Powe NR, Kohane IS, Manrai AK. Clinical Implications of Removing Race From Estimates of Kidney Function. JAMA. 2020 PMID: 33263721 and newly published manuscript Diao et al. https://jasn.asnjournals.org/content/early/2022/11/10/ASN.2022070818	We address the nuances of removing race and ethnicity as an input variable in the Results, but sought to present a concise overview in the Introduction, necessarily limiting the nuance presented.
Peer Reviewer #1	Introduction	Page 1 lines 24-28 the inclusion of R/E variables are exactly what is needed to determine if there are biases against the various R/E groups.	Data about race and ethnicity are needed to conduct such an assessment, but such data do not need to be included in a model's training process as an input variable.
Peer Reviewer #1	Introduction	Page 1 lines 29-36 This is one study – there are numerous other studies that delineate how an algorithm contributes to biases, especially the harm due to minorities.	We agree, but highlight the Vyas article given its substantial influence on focusing policymakers on the concerns addressed in this review.
Peer Reviewer #2	Introduction	The introduction sets an appropriate frame for the paper, summarizing the important analysis of Vyas et al. (page 1), and summarizing the discredited rationale for race-based clinical algorithms. Page 1, lines 37-48, are particularly good in this respect and set the tone for the paper.	Thank you for this acknowledgement.
Peer Reviewer #1	Introduction	Page 1 39-48 This paragraph is bases on one perspective. There are numerous articles that identify biological differences that contribute to racial/ethnic disparities in disease and disorders. Articles include other issues include drug impacts, for instance Warfin is not effective in African Americans. This paragraph needs to be written more objectively and should capture health outcomes of racial and ethnic groups that lead to the very large field of study regarding health disparities.	We acknowledge that differing perspectives exist, and that race and ethnicity are important constructs in health, healthcare, and research on disparities. We have made some edits to the text for clarity. However, we explicitly embrace our stated perspective as an underlying premise of this review, in consultation with our Subject Matter Experts, Key Informants, Technical Expert Panel, and AHRQ.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research

Published Online: December 8, 2023

Commenter	Section	Comment	Response
TEP/KI Reviewer #1	Introduction	Line 54 “Furthermore, exclusive categories do not capture multi-racial and ethnic individuals”, some of which been present for centuries (e.g., average mixed African, ~75-80% and European,~20-25%, ancestry of descendants of American slaves).	Thank you for providing this context.
Peer Reviewer #1	Introduction	Page 1 – Line 54 this is not correct OMB has a mixed-race option and so does the census and many other instruments. The references are outdated 2009 and should be updated to include the current standards.	This paragraph was revised for clarity and accuracy.
Peer Reviewer #1	Introduction	Page2 line 5-6 This is an misstatement. The variability comes from local systems or researchers wanting to capture their population groups more granularly so they add on to the standards. Real Standards are not what is used because it is recommends from NAM. The standards are set by OMB.	This paragraph was revised for clarity and accuracy.
TEP/KI Reviewer #1	Introduction	“Developers of healthcare algorithms sometime justify the inclusion of racial/ethnic input variables by citing observational studies or post hoc analyses of trial data that demonstrate differences in characteristics or outcomes among different racial/ethnic groups. These studies may be small and unrepresentative, serve to reinforce misconceptions, or assign race/ethnicity as a contributing cause when other factors may be causative, confounding, or modifying the effects of race/ethnicity.[20,21] A robust example in the published literature examines a “race-correction” coefficient in creatinine based equations that raises the threshold of concern for Black patients only for a given of the estimate glomerular filtration rate (eGFR), a key indicator in diagnosing and treating kidney disease.” This built on previous national data from NHANES demonstrating that creatinine levels at every age are higher in Black men and Black women compared to their White counterparts, questioning the use of a single threshold for both racial groups. (see Jones C 1998 AJKD AND Powe N 2002 Med) Reference: Jones CA, McQuillan GM, Kusek JW, Eberhardt MS, Herman WH, Coresh J, Salive M, Jones CP, Agodoa LY. Serum creatinine levels in the US population: third National Health and Nutrition Examination Survey. Am J Kidney Dis. 1998 Dec;32(6):992-9. Note that Camara Jones, notable racism scholar, was an author of this study and her sister Camille was first author) AND Powe NR. Race and Kidney Function: The Facts and Fix Amidst the Fuss, Fuzziness and Fiction. Med (Cell Press) 2022 3: 93-97	Thank you for providing these references, which we might include in the final report.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
TEP/KI Reviewer #1	Introduction	Recent studies have modeled the effect of removing the race-based coefficient22-24 and concluded that while Black patients might receive needed kidney transplants earlier but might lead to underuse or underdosing of important medications, less eligibility for living kidney donation and less enrollment of Blacks in clinical trials without dropping the race coefficient **Medications Reference:** Casal MA, Ivy SP, Beumer JH, Nolin TD. Effect of removing race from GFR-estimating equations on anticancer drug dosing and eligibility: a retrospective analysis of NCI phase 1 clinical trial participants. Lancet Oncol. 2021 ;22(9):1333-1340. And Duggal V, Thomas IC, Montez-Rath ME, Chertow GM, Kurella Tamura M. National Estimates of CKD Prevalence and Potential Impact of Estimating Glomerular Filtration Rate Without Race. J Am Soc Nephrol. 2021 May 6: **Trials Reference:** Charytan D, Yu J, Jardine M, Cannon C, Agarwal R, Bakris G, Greene T, Levin A, Pollock C, Powe N, Arnott C, Mahaffey K. Potential Effects of Elimination of the Black Race Coefficient in eGFR Calculations in the CREDENCE Trial. Clin J Am Soc Nephrol. 2022 Jan 21:CJN.08980621. doi: 10.2215/CJN.08980621. PMID: 35063969.	Thank you, we might add some of these references in the final report.
Peer Reviewer #1	Introduction	Page 2 line 24 They thought it best because of another input variable that countered the race variable, which was the variable that prioritizes those who donated a kidney prior – that population group is white women.	Thank you for this context, we are revising the description of the Task Force's recommendations for clarity.
TEP/KI Reviewer #1	Introduction	However, controversy around this issue remained,25-27 as the evidence base lacks prospective trials comparing differing approaches to assessing kidney disease and subsequent need for treatments including transplant. Accordingly, the National Kidney Foundation and the American Society of Nephrology convened a task force to address this topic. In September 2021,the task force released its final report recommending 28: immediate implementation of 2021 CKD-EPI creatinine equation refit without the race variable in U.S. labs and national efforts to facilitate increased, routine; timely use of race-independent lab-based biomarker cystatin C, for confirmation of eGFR calculations with creatinine and investment in science on new GFR markers and intervention to eliminate racial and ethnic disparities. Comment: Note, As written this was wrong. The Task Force did NOT recommend replacement with cystatin C. This was for confirmation.	We have clarified this point.

Source: *https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research*

Published Online: December 8, 2023

Commenter	Section	Comment	Response
TEP/KI Reviewer #1	Introduction	Line 25-47 are very important and insightful paragraphs. Note that the NKF and ASN instructed its Task Force to eliminate race after the deliberations started due to pressure by advocates. Fortunately, the Task Force found an evidence-based way to do this, not simply by dropping the race coefficient from the existing equation.	Thank you for providing this valuable, behind-the-scenes context.
Peer Reviewer #1	Introduction	Page 3 line 21 should have a comma after healthcare before such	Thank you, this correction was made.
Peer Reviewer #1	Introduction	Page4 lines 10-13 Please explain how you ascertained these algorithms to test them and on what data sets?	We describe the process of selecting and evaluating these algorithms in the Methods, and we also revised this paragraph in the Introduction for clarity.
TEP/KI Reviewer #1	Methods	Line 21 January 1, 2011, to January 12, 2022: The literature in kidney disease disparities in transplant and nephrology referral dates back to 1980's and early 1990s, and differences in creatinine were noted by Jones et al as mentioned above in 1988, one year before the first racebased eGFR equation was published in 1999 (Levey et al). The review period therefore misses critical timing and information with regard to eGFR.	Based on guidance from Subject Matter Experts, Key Informants, and our Technical Expert Panel, articles published before 2011 were unlikely to be directly relevant to algorithms currently in use. Although we agree that earlier studies can provide important background and context, the very large volume of studies we initially identified for potential inclusion necessitated that we limit our review to more recent research in order to satisfy project timelines.
Peer Reviewer #2	Methods	The exclusion criteria would be appropriate for a study focused on benefits but are too stringent for a study focused on harms. The exclusion criteria skew the results and conclusions towards saying the evidence is too limited to speak to harm from currently used algorithms. The choice to include many algorithms with potential disparate impact regardless of whether they had an explicit race input variable led to a less comprehensive assessment of the impact of including a race input variable (the authors noted in limitations there were too many studies to include in the chosen broad approach).	Thank you for this critique. We have now expanded both our summary, introduction, and discussion sections to ensure we clearly state the project scope and the limitations and implications of the broader scope when it comes to discussing harm. We will also delineate any potential patterns found (if at all) when it comes to causing harm.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #2	Methods	The response to the key question about harm should include using population level data on prevalence to estimate the range of potential harm based on currently available evidence about how widespread use of the algorithm is, disaggregated by race and ethnicity. The scale of the problem is obscured by the current approach.	Thank you for this critique. We have now expanded both our summary, introduction, and discussion sections to ensure we clearly state the project scope and the limitations and implications of the broader scope when it comes to discussing harm. We will also delineate any potential patterns found (if at all) when it comes to causing harm.
TEP/KI Reviewer #1	Methods	Did you look at how race/ethnicity was captured in the studies? This was discussed in the methods, but were you able to capture this in the studies? Was it self-reported or inferred by the healthcare system. This is an important difference. There should be a comment about this in the limitations section	We did attempt to determine, when available, how the studies captured data on race and ethnicity, and reported this in the Results. We will consider expanding on the potential impact of this limitation in the Discussion, during the public comment period.
TEP/KI Reviewer #1	Methods	What about mixed race - the fastest growing category in the 2020 census? There should be a more discussion on mixed race category in the discussion	Thank you for this suggestion. We will consider how to address this issue during the public comment period.
TEP/KI Reviewer #2	Methods	As I stated above, I as there is not an abundance of research in this area, I would have appreciated a summary of the findings from non-US studies and those that did not have outcomes measurements. These may have provided some additional frames to consider.	As above, the scope and timeline of this project made it infeasible to review and summarize such studies, even in a very general manner.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Methods	No, the methodology is unclear and not sufficient to assess the Key Questions or the Content Questions. The methodology section is a restatement of the exec summary and preface. How did they get the data sets from publications to do lines 42-47? They cannot do such without the raw data. There is no mention of the disparate data sets, especially between whites and blacks, as well as other racial/ethnic groups. This disparity impacts the validity of the algorithm and is addressed often through the use of synthetic data. This is also not mentioned. Implementation identification of biases can be determined by other means than empirical data. There were articles on images in algorithmic biases. There are many ways to test for biases…what methods did they use regarding the testing and retesting of inputs and with/without race/ethnicity—using one method, such as ROB, is not sufficient. They did not include the methods used to determine why a variable should be input or extracted, such as random forest etc. CQ1-3 and 4 are statements not a description of the methodologies used. This section needs to be reworked so one can determine the methodology of the project reviewed.	We have revised some sections of the Methods for clarity. We note that this review was not intended to examine raw data sets used to derive or validate the existing algorithms in Key Question 1, and studies on use of images were excluded in consultation with AHRQ.
Peer Reviewer #1	Methods	Page 8 lines 21-25 They did not include technology focused journals	We reviewed databases and journals that we expected to be most likely to publish studies that met our inclusion criteria, in consultation with our Subject Matter Experts, Key Informants, and Technical Expert Panel (which included experts in health technology). We recognize that other sources might include relevant research, and we anticipated that any key studies we missed would be identified by the experts mentioned above, or from Peer Reviewers and those who provide public comment.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Methods	Page 10 line 10 This is a fatal flaw of the literature review. Empirical studies may not reflect the disparate issues in algorithms. The algorithm biases are often found before empirical studies are conducted and most important most AI applications DO NOT DO EMPIRICAL STUDIES BEFORE IMPLEMENTING	We agree that algorithms are often not tested empirically before implementation. Nevertheless, the primary intent of Key Question 1, in consultation with AHRQ, was to identify studies that presented empirical evidence of how algorithms affect clinical outcomes. Although we anticipated that such studies would be a small subset of all available research on this topic, these studies likely present the most relevant and credible results for key stakeholders.
Peer Reviewer #1	Methods	Page 10 line 19 This is another fatal flaw. The biases are not typically noted so aren't corrected before implementation. By only using those that have a mitigation strategy, the review is limited to typically one attempt to correct the algorithm. This is typically not achievable in a single strategy.	The intent of Key Question 2 was to summarize evidence on the effectiveness of mitigation strategies. Therefore, we included studies if they applied a mitigation approach (or multiple approaches) to an existing algorithm, and then measured the results.
Peer Reviewer #1	Methods	Page 10 Line 30 This is another fatal flaw. If the algorithm didn't have a r/e variable how was it determined if it had biases or not towards r/e groups? These biases would not be determined until application, which would then make the article excluded because it is not an empirical study.	We included studies that examined if racial and ethnic disparities were evident after application of an algorithm to a patient population. Disparities in outcomes could exist even when race and ethnicity were not input variables. We acknowledge, however, that establishing causality in such cases can very difficult, and discuss this challenge in the report.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Methods	Page 10 Line 46 This is another fatal flaw. This narrows the evaluation to minimal. This type of approach is rarely done in the data science field. This is not a clinical trial.	We acknowledge that restricting the review to studies that actually reported results on racial and ethnic differences (or lack of differences) limits the evidence we could review. However, after extensive consultation with AHRQ and our Subject Matter Experts, Key Informants, and Technical Expert Panel, we concluded that these types of studies were best suited for an evidence base that could inform action by key stakeholders.
Peer Reviewer #1	Methods	Page 10 Line 51 this is another fatal flaw. Data science isn't looking for effect sizes. In unstructured data, the algorithm teaches itself	We use "effect size" to refer to the degree of difference in model performance between subgroups, which are not infrequently reported in studies using data science methods. Unstructured data typically refers to text data, which was not addressed here, and studies of unsupervised learning methods were not included in this review.
TEP/KI Reviewer #1	Methods	Page 26 Table 1. For eGFR another important outcome is access to clinical trials as shown in Charytan D, Yu J, Jardine M, Cannon C, Agarwal R, Bakris G, Greene T, Levin A, Pollock C, Powe N, Arnott C, Mahaffey K. Potential Effects of Elimination of the Black Race Coefficient in eGFR Calculations in the CREDENCE Trial. Clin J Am Soc Nephrol. 2022 Jan 21:CJN.08980621. doi: 10.2215/CJN.08980621. PMID: 35063969.	Thank you for highlighting this important point. Insofar as access to trials is rarely reported as a quantified outcome within studies, we did not address it directly. But we will consider how this point might be included in our Discussion.
TEP/KI Reviewer #1	Methods	Where should opportunity to be a living donor for kidney transplants be. This is important in the eGFR outcomes. Research might be added to the Setting category under non clinical sites.	Thank you for this suggestion.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Methods	Page 11 line 3-6 If this is only clinical applications, then the scope is too narrow. It is also another fatal flaw if synthetic data was not included. There is no mention of synthetic data or digital twins, which is the core issue of algorithm biases in clinical trials.	Synthetic data and digital twin studies are not widely used in the development and evaluation of clinical prediction models, or in the evaluation of model performance by demographic subgroups; therefore, we did not address them in this report.
Peer Reviewer #1	Methods	Page 12 line 28 needs a coma after complexity before we	This revision was made.
Peer Reviewer #1	Methods	P12 line 34 – Before an adapted instrument/algorithm can be used, it has to be tested/validated. ROBINs was adapted.	We revised this section to improve clarity of our approach.
Peer Reviewer #1	Methods	Page 13 line 16 –missing data is one of the major causes of algorithm biases – restricting its inclusion is a fatal flaw.	We agree that missing data is a major contributor to bias. However, our risk of bias assessment as applied to Key Question 1 focused on how a given study of a previously validated algorithm reported data; we were not evaluating, for this purpose, whether earlier derivation and validation was conducted with optimal data. We recognize that this is a limitation of the review.
Peer Reviewer #1	Methods	Page 13 line 24-25 is critical to understanding and identifying biases and health disparities. This reflects a lack of understanding of the impacts.	We revised this section to improve clarity.
Peer Reviewer #1	Methods	Page 15 Diagram - How is implicit bias or explicit bias being measured? This diagram looks good on paper, but it is not operationalized. Literature is not going to detect bias. They could detect difference.	This diagram was designed to address Contextual Question 4, which aimed to examine the characteristics of a sample of algorithms previously unstudied for possible racial and ethnic bias.
TEP/KI Reviewer #1	Methods	In Table 2 Lack of representation of racial and ethnic minorities/ selection in the dataset. This was a big issue in the Cockcroft Gault equation developed in 1976 and recognized in the MDRD and CKD Epi equations for eGFR. The removal of the race -coefficient used in many of the simulation studies for eGFR is akin to removal of the creatinine data on Black persons and has no evidentiary basis compared to the refit equation developed by the NKFASN Task Force.	Thank you for providing this important context.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Methods	Page 16 line 42-43 This is not a measure of bias—either implicit or explicit	We are unsure what this comment means to convey, but the concepts described in this paragraph are not intended to represent measures of bias. They are identifying key entry points and mechanisms by which biases can become relevant during implementation.
TEP/KI Reviewer #1	Methods	Page 31 line 44. Please realize that nearly all African-Americans who are descendents of slaves are “mixed-race” in terms of genetic ancestry not just someone whose immediate parents are of different races.	Thank you for your comment.
Peer Reviewer #1	Methods	Page 17 line 7-10 Please remove all the “right” in this system – many algorithms are implemented without knowing if any of these are right.	We agree that algorithms are often implemented without a clear understanding of the relevant factors, but this sentence is intended simply to convey the ideal goal of implementation.
Peer Reviewer #1	Methods	Page 17 line 11-14 There is no validity or merit to this sentence because there is no proof or measurement of a clinician’s bias. Clinicians implement the algorithm because they assume it has been tested and validated. The biases that result are a result of the algorithm and the lack of validation before implementing. It is not exacerbated by the clinician’s biases.	We disagree with this point based on our own experience implementing algorithms within an academic medical center. Clinicians often exercise personal discretion in how they use, interpret and apply the results of algorithms. Clinician biases, both implicit and explicit, can add to algorithmic biases and exacerbate the potential harm for patients. Similarly, algorithmic biases can reinforce biased beliefs that a clinician may have about a given patient population.
Peer Reviewer #1	Methods	Page 17 lines 48-51 The experts could only rule upon what was given, which was a narrow scope at a stage where little to no biases could be detected.	We have added additional details to ensure we clearly state the scope of our work and the limitations and implications of that scope.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Methods	Page 18 Line 17-18 Adding additional variables changes the derivatives.	The additional variables referred to in this sentence are not algorithmic variables used within the studies, but instead refers to characteristics of the studies that our team chose to abstract in order to describe the studies. We revised this sentence for clarity.
TEP/KI Reviewer #1	Results	Note well that the original CRIC investigators reanalyzed the CRIC data used in Reference 72 by Zelnick LR, Leca N, Young B, et al. Association of the estimated glomerular filtration rate with vs without a coefficient for race with time to eligibility for kidney transplant. JAMA Netw Open. 2021Jan;4(1):e2034004. doi:10.1001/jamanetworkopen.2020.34004.PMID: 33443583. They found a fundamental analytic flaw published here. Hsu C, Yang W, Go AS, Parikh RV, Feldman HI. Analysis of Estimated and Measured Glomerular Filtration Rates and the CKD-EPI Equation Race Coefficient in the Chronic Renal Insufficiency Cohort Study. JAMA Netw Open. 2021;4(7):e2117080. doi:10.1001/jamanetworkopen.2021.17080. They concluded "we do not believe CRIC data support the notion that dropping the race coefficient in the current CKD-EPI equation enhances accuracy in kidney function estimation." The Zelnick paper might be excluded or this correct reanalysis by the experienced CRIC investigators mentioned alongside.	Thank you for identifying this issue. We will reconsider our inclusion and/or evaluation of the Zelnick study during the public comment period.
Peer Reviewer #2	Results	The results are appropriately detailed and clearly described. As stated above, too many studies were excluded due to overly stringent criteria for a study on harms, and the study lacked comprehensiveness on algorithms with an explicit race input due to the inclusion of other algorithms that may have disparate impact, so this impacted the results. In addition, there are no results on the potential scope of the problem given current or recent population prevalence and use of these tools (for example, the indications of "high" or "moderate" prevalence in Table 8 are inadequate).	Thank you.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Results	The details of the results and the key messages are not clear and some of the assumptions of the meaning of the analysis results are not applicable. The scope covered in inadequate because of the limited resource they scanned for literature. Figure 1 is inadequate and overstates what they actually did in the analyses. There were only a few disease/disorders covered. Clinical info was assumed to be the only articles used; however, some were not clinical. The results don't align with Figure 1 or Table 1. Investigators overlooked many studies. Most of the testing of an algorithm is not published in pubmed or medline. I am just listing a few of the many articles not reviewed that are critical. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research Y Juhn, H Liu - Journal of Allergy and Clinical Immunology, 2020 – Elsevier Exploratory Study of Artificial Intelligence in Healthcare R Alugubelli - International Journal of Innovations in Engineering …, 2016 - academia.edu Bias, Fairness, and Accountability with AI and ML Algorithms N Zhou, Z Zhang, VN Nair, H Singhal, J Chen… - arXiv preprint arXiv …, 2021 - arxiv.org A survey on bias and fairness in machine learning N Mehrabi, F Morstatter, N Saxena, K Lerman… - ACM Computing …, 2021 - dl.acm.org Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. CG Walsh, B Chaudhry, P Dua, KW Goodman… - JAMIA …, 2020 - academic.oup.com Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index YJ Juhn, E Ryu, CI Wi, KS King, M Malik… - Journal of the …, 2022 - academic.oup.com An individual-level socioeconomic measure for assessing algorithmic bias in health care settings: A case for HOUSES index. YJ Juhn, E Ryu, CI Wi, KS King, SR Brufau, C Weng… - medRxiv, 2021 - medrxiv.org Can AI be racist? Color-evasiveness in the application of machine learning to science assessments T Cheuk - Science Education, 2021 - Wiley Online Library The practical implementation of artificial intelligence technologies in medicine J He, SL Baxter, J Xu, J Xu, X Zhou, K Zhang - Nature medicine, 2019 - nature.com Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types D Rozado - PloS one, 2020 - journals.plos.org	We have substantially revised the Results for clarity. We appreciate the reviewer identifying potentially relevant articles, and address each one below. Most of them are narrative reviews, several are not specific to healthcare, and none present data that directly address our Key Questions. Juhn 2020: natural language processing was outside the scope of this review. Alugubelli 2016: this is a narrative review. Zhou 2021: this is a narrative review. Mehrabi 2021: this narrative review focuses on technical types of bias and is not specific to healthcare. Walsh 2020: this is a commentary. Juhn 2022 and 2021: these studies examine the link between socioeconomic factors and algorithmic bias, but do not address our Key Questions. Cheuk 2021: this is a commentary. He 2019: this is a narrative review. Rozado 2020: does not examine a clinical algorithm.
TEP/KI Reviewer #1	Results	The results included sufficient detail	Thank you for your review.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
TEP/KI Reviewer #2	Results	The results are well described and comprehensive. I liked the focus both on the issues with algorithmic bias as well as the mitigation strategies.	Thank you for your review.
Peer Reviewer #2	Results	An additional minor query pertains to the exclusion of papers that were "not a full-length article" (p. 35, line16). Additional details on this would be helpful, as it is unclear if a brief research letter/brief would be counted as a "full-length article". Given that this took out n=23 and there is n=46 in the final set included in the analysis, the distinction could be important.	The exclusion category includes reviews and commentaries that do not present original data.
TEP/KI Reviewer #1	Results	Table 3. It is very important to specify whether race was removed in the calculation versus removed in the reporting as well as which race was removed. Many of the eGFR studies are ones where Black race was removed in the reporting and Black persons were assigned the White value rather than the other way around. The Inker et al 2021 removed race by refiting the equation, a new calculation that is evidence based unlike removal of race in reporting which is not calculation based and creates predicition biases. The big point when making these comparisons is "how is race removed?".	Thank you for providing this context. We will continue revising our treatment of the eGFR algorithm to ensure accuracy and clarity.
Peer Reviewer #1	Results	PAGE 26 Line 4-17 Confounding domains are a separate issue than race/ethnic biases – all of the elements listed and more jeopardize the finding the authors are reporting.	In our risk-of-bias evaluation, bias due to confounding is one of seven domains assessed. We agree there are several areas of concern that threaten the validity of the causal impact of healthcare algorithms on racial and ethnic disparities in outcomes.
Peer Reviewer #1	Results	Page 29 line 53 to Page 30 line 3 If the white patients were older, they will have higher mortality rates. Also, the critical factors being assessed are not the health areas that most impact African Americans mortality, such as stroke, kidney failure, diabetes, sepsis, etc.	We agree with both of the reviewer's points, but in this paragraph we are simply describing SOFA scores as evaluated in the included studies.
Peer Reviewer #1	Results	Page 31 line 12-13 Explains the flaw. There were 16K Black patients and 95K white patients. The model needs more data on Blacks.	While the sample of White patients was certainly much larger than that of Black patients, the required sample size to fit underlying models depends on the event rate, the degrees of freedom of the model, and many other factors that were not reported.
Peer Reviewer #1	Results	Page 31 line 32-33 The significance of the lower risk for death in these models are likely due to the death risk factors included in the model. The measures in all three approaches are more aligned with white mortality factors than other racial/ethnic groups.	Thank you for your comment.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Results	Page 32 line 7-8 Again the disparity in sample size impacts the validity of any algorithm to accurately performs-88% are white and 12% are Black. This is a consistent fatal flaw and should be noted as a reason for the lack of model accuracy in algorithm performance.	We agree that poor representation of patients who are Black (or other non-White races and ethnicities) is often a key flaw in algorithm development, and address this in the Discussion.
Peer Reviewer #1	Results	Page 32 line 24-25 These authors missed the real discrepancy of this study that was revised. Costs was associated with the number of office visits. Number of office visits was used in the algorithm and used to indicate need. The real issue was Blacks are less insured and work in jobs more likely not to have sick leave so make fewer office visits; thus, creating the flaw in the algorithm.	We reported that the authors found that, at any given level of health, Black patients generate less costs than White patients. This phenomenon is likely explained by a number of different factors, but it was generally outside the scope of this project to suggest any such factors.
TEP/KI Reviewer #1	Results	Page 47 line 30. The following studies should be added that used NHANES and are much more representative than other limited setting studies from one institution. See Diao JA, Wu GJ, Taylor HA, Tucker JK, Powe NR, Kohane IS, Manrai AK. Clinical Implications of Removing Race From Estimates of Kidney Function. JAMA. 2020 PMID: 33263721 and newly published manuscript Diao et al. https://jasn.asnjournals.org/content/early/2022/11/10/ASN.2022070818	Thank you for suggesting these suggested references. Diao et al. 2020 was identified in our literature searches and cited in the introduction. Diao et al. 2022 was published after the initial literature search. We will update our search and review this paper during the public comment period.
TEP/KI Reviewer #1	Results	Page 47 Line 39-43. The Task Force recommendations are not correctly stated. As previously mentioned, in September 2021, the National Kidney Foundation/American Society of Nephrology (NKF/ASN) task force released its final report recommending discontinuing the race variable in calculating eGFR with creatinine and ~~replacing it with a lab-based~~ increased use of the biomarker, cystatin C for confirmation of eGFR calculated with creatinine; a revised version of the CKD-EPI equation has subsequently been developed	This section was substantially revised and the description of the recommendations were removed.
TEP/KI Reviewer #1	Results	Line 47-49 should read To modify equations, all three studies removed the race variable from reporting (not the calculation), and one[58] added body surface area to a Deindexed CKDEPI.	This section was substantially revised and the sentence removed.
Peer Reviewer #1	Results	Page 32 line 53-55 AGAIN, the disparate number of other racial/ethnic groups creates a challenge for algorithms to function properly. This is such an issue that AI developed synthetic data.	We agree there is need for more diversity in algorithm derivation and validation populations, and address this in the Discussion.

Commenter	Section	Comment	Response
Peer Reviewer #1	Results	Page 33 line 32-33 Is a judgement statement made in error. The other r/e groups did not have a sufficient to even be included. This sentence should be removed.	After further review, we removed this study from Key Question 1, and only discussed its results as relevant to Key Question 2. The sentence highlighted by the reviewer was removed.
Peer Reviewer #1	Results	Page 34 line 36-37 This is an important component of this analysis. The algorithm alone may not have been the deciding factor. This study also gave preference to people who donated a kidney, as appears appropriate. This also created a bias because those most likely to donate kidneys are white women; thereby, increasing the preference for whites. Removing the race variable made this solely a gender issue.	After further review, we removed this study from Key Question 1, and only discussed its results as relevant to Key Question 2. The findings referred to in this comment are no longer addressed in our review.
Peer Reviewer #1	Results	Page 36 Paragraph that starts with line 15 is a statistical analyses, not an algorithm. The last lines of paragraph depict the differences, as the white patients are older with more co-morbidities, which is likely because kidney failure is a health disparity for African Americans.	After further review, we determined this study did not meet eligibility criteria (i.e., does not examine a clinical algorithm or algorithm-based tool) and have excluded the study from our evidence base.
Peer Reviewer #1	Results	Page 37 line 35 indicates vastly different sample sizes of whites and blacks; thereby, influencing the results. Page 38 Line 23-27 indicates disparate sample sizes which will impact results. Page 39 line 15-16 indicates disparate sample sizes which will impact results.	We agree there is need for more diversity in algorithm derivation and validation populations, and address this in the Discussion.
Peer Reviewer #1	Results	Page 40 the Opioid use algorithm has no summary result. There should be a translation of the two correction strategies.	The summary has been revised to address this point.
Peer Reviewer #1	Results	Page 40 lines 6-7 make my prior point in the studies above that have 80-90% white and 10-20% black were ineffective algorithms. The sample size has to be more aligned	For Key Question 1, we assessed the effect of algorithms on racial and ethnic disparities and when summarizing studies, were constrained to data reported. We agree there is need for more diversity in algorithm derivation and validation populations.
Peer Reviewer #1	Results	Page 41 line 26 – what are the results for the first question?	We are unsure to which question this comment refers.
Peer Reviewer #1	Results	Page 41 line 54-55 appears to be a study testing the ability of the algorithm to predict better than a biopsy – Intent of the study is critical to note.	The summary has been revised. The first paragraph now includes the intent of the studies.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
TEP/KI Reviewer #1	Results	Page 61 line 17 There were many more outcomes of changes in chemotherapy in the study by Casal that this one, some of which were more profound---Casal et al.[50] reported that 26% of Black patients (90/340) who were undergoing cancer treatment were reclassified as having a more severe kidney disease after the race coefficient was dropped; however, 5% (18/340) were newly deemed ineligible to receive cisplatin after the removal of the race coefficient because their revised renal function estimate exceeded standard medication safety thresholds.	These specific outcomes were selected because we felt they were most consistent with our inclusion criteria and with the types of outcomes reported in other studies we included.
Peer Reviewer #1	Results	Page 48 line 3-6 In opposition to the background that states race/ethnicity is a social construct, then adding biomarkers should not increase accurate risk predictions for patients of all racial, ethnic and ancestral backgrounds.	In this sentence we are reporting the intention of the study's authors. We acknowledge that some, and perhaps many, of the studies included in this review are not premised on the recognition that race and ethnicity are socially constructed.
TEP/KI Reviewer #1	Results	Page 63. An algorithm's components and construct are affected substantially by the characteristics of the patients used for derivation and validation. When relevant populations are not adequately represented during development or their contribution are removed by reporting of race rather than recalcuation**,** an algorithm may reflect and contribute to racial differences.	Thank you for suggesting this edit.
Peer Reviewer #2	Results	Additionally, some parts of the Discussion could use some refinement – in particular, page 55/56, where the authors write "…many if not most Americans (including BIPOC communities) do not conceptualize race as a thoroughly social construct or understand the mechanisms of structural racism." It would be helpful to provide references for that statement, which may or may not be grounded in empirical data. Similarly, on page 56, there is this claim: "Recent controversies about eGFR and other algorithms may have attracted broad attention but do not seem to have significantly affected public opinion or patient awareness." There is more to unpack in that statement (including the evidence for it as well as an analysis of the expectation that general public opinion could be captured by health care algorithms).	We have revised those statements to clarify that our conclusions are based on our discussions with the Key Informants and Technical Expert Panel.
Peer Reviewer #2	Results	Table 9, detailing variability in the definitions of race and ethnicity used in these studies, is particularly good, detailing that "algorithm developers did not specify race and ethnicity definitions, nor were race and ethnicity consistent with available standards for race categories." (p. 80)	Thank you for your review.

Commenter	Section	Comment	Response
TEP/KI Reviewer #1	Discussion/ Conclusion	Another point to mention and emphasize is that most removal of race in reporting always uses the White or majority as the standard. This happened in genonomics field as well. Implementation of removal or race could have used the Black value as the standard (rather than assigning the white value at the best value to use) which likely would have more saluatory effects for Black persons.	Thank you for emphasizing this point. We are considering how to address this issue in the final report.
Peer Reviewer #2	Discussion/ Conclusion	The implications are clearly stated. The limitations are adequately described. Regarding important literature, the authors made a decision to broaden to algorithms without an explicit race input variable and then said there were far too many studies to do a comprehensive review, so it should perhaps be acknowledged that this was a strategic error and the focus should have been a more comprehensive review of algorithms with an explicit race input variable.	Thank you for your thoughtful review of our report. We note that the decision to include algorithms that do not use race or ethnicity as an input variable was made in consultation with AHRQ, to best serve the needs of relevant stakeholders, and with feedback from our Subject Matter Experts, Key Informants, and Technical Expert Panel.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023

Commenter	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	Implications of the findings are poorly stated with little foundation to make some assumptions. The assumptions made often reflect a lack of knowledge of data science and/or medicine. Cut and pasting portions of an article does not advance the understanding of the biases in algorithms. The last sections have little to know references and justifications for what was stated. The conclusions mostly do not support that there are biases in algorithms and the ones initially identified are mitigated by 6 strategies. The six corrective strategies are inadequate and only represent a small set of potential corrective actions. The strategies used in Python or R were not mentioned which is key to algorithm bias detection. This is not true!! Missing data and disparate sample sizes are not mitigated by any of these strategies; thus, showing their lack of knowledge. Clinical use of AI relies heavily on synthetic and digital twins, which was not mentioned at all. The omitted very important literature because most of the testing of an algorithm is not published in pubmed or medline. Many of the articles in their references are statistical and algorithms used in AI, which is the intent of the Congressional request.	We are grateful for the comprehensive review, honest critique, and specific feedback that have helped improve the report. We note that the individuals who drafted, critically revised, and thoroughly reviewed the report prior to peer review include, among others, six practicing clinicians, three experts in data science and the development, evaluation, and implementation of clinical algorithms, and four experts in health equity and disparities research. We have revised the Discussion and will continue to make revisions during the public comment period. We agree that this report does not address certain types of mitigation strategies, and we did not examine tools that are used to detect (rather than proactively mitigate) bias in clinical algorithms. However, the mitigation strategies described in the report do include approaches to dealing with poorly representative data sets. Meanwhile, synthetic data and digital twins, while undoubtedly significant in artificial intelligence applications, are not generally incorporated into the development or implementation of the types of clinical prediction models studied in this report. Finally, our literature searches were more expansive than the standard clinical literature databases, and included resources specific to data science, computing, and informatics.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research

Published Online: December 8, 2023

Commenter	Section	Comment	Response
TEP/KI Reviewer #1	Discussion/ Conclusion	It would be great to have a bigger discussion on the evidence gaps. These are important issues and the community as a whole are struggling with these gaps. Is there a call to action for Congress?	Thank you for your suggestion. We have added an evidence gaps section and recommendations for key stakeholders.
TEP/KI Reviewer #1	Discussion/ Conclusion	The FDA now has a Predetermined change control plan, perhaps you can mention this in the discussion	We are examining this plan and will determine whether to reference it in the final report.
TEP/KI Reviewer #2	Discussion/ Conclusion	The authors are conservative on their speculation of the impacts of use of race/ethnicity in algorithms. As they stated, kidney and lung transplantation algorithms prioritize more people of color after removing race from the data used on the waiting list. This is a big deal. These are the algorithms that have published analyses- there may be many others with as much or more disproportionate allocation of resources due to algorithms containing race.	We have modified the Results and Discussion to emphasize the impact a little more strongly, but we are also constrained by the evidence we have identified.
TEP/KI Reviewer #2	Discussion/ Conclusion	I would like to see a stronger set of recommendations for future research. The authors certainly have a number of hypotheses that they now feel should be tested and methods the would like to see used. I would like a longer explication.	Thank you for your suggestion. We have added a series of recommendations at the end of the report.
Peer Reviewer #1	Discussion/ Conclusion	Page 80 line 13-14 although the statement is true, there was no mention of racism and ways to measure such	Thank you for this feedback.
TEP/KI Reviewer #1	Discussion	Page 96, Line 19 NHANES should be included with the Diao et all studies. This is one of the most representative data sets in the U.S. Included studies frequently used national data (e.g., NHIS, CISNET, US Census, United Network for Organ Sharing waitlist, eICU Collaborative Research Database, MIMIC-III) rather than data from a few local hospitals, increasing applicability of findings to broad populations. National datasets provide a more representative distribution of races, algorithms typically used widely available input variables (and thresholds), and the levels of access to care and health outcomes more accurately reflect the United States as a whole.	We will determine whether to include the recent Diao study during public comment, and will also consider how we might incorporate the value of national data sets in the Discussion.
TEP/KI Reviewer #1	Discussion/ Conclusion	Line 26-32. This is a very important point. The CKD Epi development did look at other ethnicities and found differences in eGFR for other race/ethnicities. See Stevens LA, Claybon MA, Schmid CH, Chen J, Horio M, Imai E, Nelson RG, Van Deventer M, Wang HY, Zuo L, Zhang YL, Levey AS. Evaluation of the Chronic Kidney Disease Epidemiology Collaboration equation for estimating the glomerular filtration rate in multiple ethnicities. Kidney Int. 2011 Mar;79(5):555-62. doi: 10.1038/ki.2010.462. Epub 2010 Nov 24. PMID: 21107446; PMCID: PMC4220293.	Thank you for making this point. We identified this study in our literature search and excluded it because it does not report an outcome of interest.

Commenter	Section	Comment	Response
Peer Reviewer #1	Discussion/ Conclusion	Page 82 line 54-55 evaluation of risk of bias should be recognized as an important limitation.	A discussion on risk-of-bias assessment has been added to the limitations section.

Source: https://effectivehealthcare.ahrq.gov/products/racial-disparities-health-healthcare/research
Published Online: December 8, 2023