Journal of Cardiac Failure Vol. 00 No. 00 2022

Review Article

Do PRO Measures Function the Same Way for all Individuals With Heart Failure?

THERESA M. COLES, PhD,[1] LI LIN, MS,[1] KEVIN WEINFURT, PhD,[1] BRYCE B. REEVE, PhD,[1] JOHN A. SPERTUS, MD, MPH,[2] ROBERT J. MENTZ, MD,[3] ILEANA L. PIÑA, MD, MPH,[4] FRASER D. BOCELL, PhD,[4] MICHELLE E. TARVER, MD, PhD,[4] DEBRA M. HENKE, BS,[1] ANINDITA SAHA, BS,[4] BRITTANY CALDWELL, PhD,[4] AND SILVER SPRING, MD[4]

Durham, North Carolina; Missouri; and Detroit, Michigan

ABSTRACT

Women diagnosed with heart failure report worse quality of life than men on patient-reported outcome (PRO) measures. An inherent assumption of PRO measures in heart failure is that women and men interpret questions about quality of life the same way. If this is not the case, the risk then becomes that the PRO scores cannot be used for valid comparison or to combine outcomes by subgroups of the population. Inability to compare subgroups validly is a broad issue and has implications for clinical trials, and it also has specific and important implications for identifying and beginning to address health inequities. We describe this threat to validity (the psychometric term is *differential item functioning*), why it is so important in heart-failure outcomes, the research that has been conducted thus far in this area, the gaps that remain, and what we can do to avoid this threat to validity. PROs bring unique information to clinical decision making, and the validity of PRO measures is key to interpreting differences in heart failure outcomes. (*J Cardiac Fail 2022;00:1–7*)

Key Words: Heart failure, women, patient-reported outcomes, differential item functioning, psychometric.

Patient-reported outcomes (PROs) provide insight directly from patients about how they feel and function.[1] PROs have long been used in cardiology clinical trials to evaluate treatment benefits and in clinical care to inform decision making about treatment. More recently, organizations such as the American Heart Association and the European Society of Cardiology have developed statements calling for more routine use of PROs in clinical practice.[2,3] Multiple PRO measures have been developed for use specifically with individuals diagnosed with heart failure (HF),[4] including the Kansas City Cardiomyopathy Questionnaire (KCCQ),[5] the Minnesota Living with Heart Failure Questionnaire (MLHFQ)[6] and the Patient-Reported Outcomes Measurement Information System (PROMIS) Plus-HF profile measure.[7] Of these, the KCCQ and MLWHFQ have been qualified by the Federal Drug Administration (FDA) as Medical Device Development Tools,[8,9] and the KCCQ has been qualified as a clinical outcomes assessment by the FDA's Center for Drug Evaluation.[10] Generic PRO measures such as the MOS 36-Item Short-Form Health Survey[11] have also been used to evaluate outcomes for individuals diagnosed with HF. This review focuses specifically on measures

From the [1]Center for Health Measurement, Department of Population Health Sciences, Duke University, Durham, North Carolina; [2]Saint Luke's Mid America Heart Institute/University of Missouri-Kansas City, Missouri; [3]Department of Medicine, Division of Cardiology, Duke University Medical Center, Durham, North Carolina and [4]Wayne State University/Central Michigan University, Center for Devices and Radiological Health, Food and Drug Administration, Detroit, Michigan.

Manuscript received March 11, 2022; revised manuscript received May 13, 2022; revised manuscript accepted May 31, 2022.

Requests for reprints: Theresa Coles, PhD, Duke University School of Medicine, Department of Population Health Sciences, Durham, NC 27701. E-mail: theresa.coles@duke.edu

See page 5 for disclosure information.

1071-9164/$ - see front matter

https://doi.org/10.1016/j.cardfail.2022.05.017

developed for use with individuals diagnosed with HF.

An important presumption in the use of PRO measures is that patients interpret PRO questions similarly, regardless of their race, ethnicity, gender, age, or other key characteristics.[12] Without formally testing for similar interpretations of questions on a PRO measure by key population subgroups, it is possible that conclusions about differences in patient-reported symptoms and functioning may be due partially to differences in interpretation of the PRO measure rather than to true differences in the outcomes being assessed. A classic example is that men and women experiencing the same level of depression may interpret and respond to a question about "crying spells" differently—not because of true variations in depression levels, but because historical social norms in the United States may bias men to under-report crying spells.[13] Hence, social pressures may be the underlying factor influencing responses instead of the concept of interest, depression. The risk then becomes that the scores cannot be used to compare or to combine PROs by subgroups of the population. Inability to compare subgroups of the population validly is a broad issue and has implications for comparing any subgroup of the population, including in clinical trials. It also has specific and important implications for identifying and beginning to address health inequities in clinical care. If PRO questions are interpreted differently by subgroups, the impact of disease or treatment in these important population subgroups would not be measured correctly, potentially leading to worse disparities. As the scientific community continues to dig deeper to ensure equitable outcomes across the population, we need to examine how outcomes are measured and what inherent biases may be present in our measurement tools.

Differential item functioning (DIF) is the psychometric term for the phenomenon that occurs when 2 groups of patients (eg, men and women, older and younger patients) interpret and respond to a PRO item differently for reasons other than the outcome of interest.[14] A hypothetical DIF example in HF could be that men and women with the same level of physical function may interpret and respond to a question about "household chores" differently—not because of true variations in physical function, but because they were thinking about different types of household chores that have different levels of physical effort. If men think of short chores (eg, wiping down a table) and women think of chores requiring more effort (eg, vacuuming stairs, grocery shopping), then they will respond differently in terms of their functional limitations. This may result in men appearing not to have limitations and women to report more limitations (because they are engaged in more difficult tasks). Thus, their scores are not equivalent because the concept is interpreted differently between genders. Ultimately, this reduces the validity of comparing physical limitations by gender or pooling data from women and men.

DIF and Heart Failure

A number of equity and disparity concerns have been raised in populations with HF.[15–18] There exist disparities in quality of care and outcomes in individuals with HF from racial and ethnic minority groups, including higher rates of hospitalization and death and potential underuse of therapies, such as heart transplantation or left ventricular assist devices, and worse patient-reported outcomes.[19,20] There is also a role of implicit bias and structural racism as key drivers for differences in outcomes by race[19] and gender.[21] When equity is investigated from patient-reported perspectives, our conclusions are only as valid as the measures we use. Gender-specific DIF, for example, may be present in PRO items developed for individuals with HF. A number of publications provide evidence of why it is important to investigate DIF in HF by gender: (1) women report worse HF-related quality of life than men with HF, but the mechanism underlying this finding is not well understood[20,22]; (2) women have different lived experiences of HF, which could cause women to interpret questions about HF differently[23–26]; and (3) there was a historical lack of representation of women in HF studies,[27] which could have excluded some women's perspectives in the development of HF measures. Another example: race or ethnicity DIF is also important to consider evaluating due to historical lack of representation of minorities in clinical studies,[27,28] which may extend to lack of representation in PRO psychometric validation studies. Valid PRO measures would ensure that experiences of all individuals with HF would be represented and measured by the same metric, regardless of gender, race, ethnicity, or any other characteristics.

What Can we Do About DIF?

There are a number of ways to identify DIF quantitatively, including an item-response theory-based method and an ordinal logistic method.[29,30,31,32] Table 1 provides an overview of the 3 prominent DIF methods and considerations for using each method. As with other statistical analyses, results might vary depending upon the type of method used. Thus, an analysis of DIF often involves multiple quantitative methods for detecting DIF. Typically, DIF methods test a single item for DIF and use a subset of the other items (called *anchor items*) in the scale to adjust for differences in the key outcome between

Table 1. Prominent Differential Item Functioning (DIF) Methods and Considerations

Method	Considerations
Quantitative methods	
Item Response Theory (IRT)-based Wald ^{2} test	• Powerful methodology to detect 2 different types of DIF: uniform DIF (differences in the threshold parameters) and nonuniform DIF (differences in the discrimination parameters)[36]
	• The data must be appropriate for the IRT model assumptions.
	• Typically requires larger sample sizes to have sufficient numbers of participants in each group being compared to allow for estimation of model parameters
	• Requires IRT software
	• Statistically-significant DIF may not require action; magnitude of DIF can be calculated to determine if changes are needed to the patient-reported outcome (PRO) measure[37]
	• Requires at least 3 items in each PRO domain
	• Analyst can select items in PRO scale to control for differences between groups on the outcome of interest.
	• Does not adjust for other covariates in model
Ordinal Logistic Regression	• Powerful methodology to detect 2 different types of DIF: uniform DIF (differences in the threshold parameters) and nonuniform DIF (differences in the discrimination parameters)
	• Analyst can use an IRT-based score or sum/average score to account for differences between the comparison groups on the outcome of interest.
	• Any statistical software with the ability to conduct ordinal logistic regression can be used.
	• Statistically-significant DIF may not require action; magnitude of DIF can be calculated to determine whether changes are needed in the PRO measure.[38]
	• Requires at least 3 items in each PRO domain
	• Can include covariates in the model to test for DIF
Qualitative Study	• Provides insight into why DIF may be present
	• Useful in refining DIF hypotheses
	• Cannot conclusively determine whether DIF is present statistically but is a viable option when sample size or anchors are not sufficient for quantitative analyses
	• PRO measure can be developed using Classical Test Theory or IRT
	• Sample size is usually small.
	• No minimum number of PRO item anchors
	• No statistical software is required.
	• Qualitative software can facilitate thematic summaries.

the 2 groups being compared. Responses to the question being evaluated for DIF are compared across individuals with the same level of the domain of interest, such as physical functioning. If a significant difference in responses to the item between the 2 groups (eg, men and women) is found after adjusting for group differences in the outcome of interest (eg, physical functioning), then the item is deemed to exhibit DIF. If the difference is deemed to be of sufficient magnitude, the conclusion is that there is something else about how subgroups interpret and respond to the DIF items that is causing the difference in item responses. The subsequent steps are to identify what may be causes of the DIF, the impact of the DIF item on the overall PRO measure score, and how it can be addressed, which may include statistical adjustment or removing the DIF item from the PRO measure. Removing DIF items from a PRO measure is often the most practical method of addressing DIF in cases in which the PRO measure includes a sufficient number of items. For example, Weinfurt and colleagues removed items to address DIF among items in a sexual-function measure.[33] If changes are made to a measure, conclusions about group differences could change due to improved validity of the measure.

Differing interpretations of PRO items could also be examined using qualitative methods, such as in-depth interviews or focus groups (Table 1). Qualitative methods can be used to refine DIF hypotheses or investigate *why* DIF is present for specific PRO items. These studies could be conducted prior to the quantitative analysis or after the quantitative portion, depending on the psychometric evidence (eg, validity, reliability, responsiveness) available for a PRO measure within the context of use (considering the population and type of research study).[14] If researchers are developing a new PRO measure or if DIF hypotheses need to be generated, researchers could conduct a qualitative study prior to a quantitative DIF evaluation to guide analyses further. A follow-up qualitative study could then be conducted to provide context for the quantitative results. When researchers have strong DIF hypotheses for an existing PRO

measure, it may be more helpful to test the hypotheses first by conducting the quantitative analysis first and then conducting a qualitative study to provide context for the quantitative results.[34] When small sample sizes are available, or when PRO measures are short (eg, < 4 items), descriptive statistics and qualitative studies may be the only useful course of action for investigating DIF. To our knowledge, there are no examples of qualitative DIF studies in cardiology, but qualitative DIF-oriented studies have been published successfully in other areas.[34,35]

State of Current Evidence on DIF for PRO Measures Developed for Adults Diagnosed With Heart Failure

DIF evaluations have been conducted in 2 PRO measures developed for adults diagnosed with HF: Patient-Reported Outcomes Measurement Information System (PROMIS+HF) and MLHFQ.

For the PROMIS+HF measure, authors evaluated sex, age and education DIF for domains with 4 or more items.[7] Of the domains that could be evaluated (≥ 4 items), no PROMIS+HF items exhibited statistically significant DIF. The domains evaluated included health-behavior outcomes, pain interference, symptoms, anger, cognitive abilities, cognitive functioning, life satisfaction, independence, and social isolation.

The MLHFQ was evaluated for gender and age (≤ 65 vs > 65) DIF in a small sample.[39] Statistically significant gender DIF was identified for 4 items. The first item asked patients about HF preventing them from living as they wanted when "walking about or climbing stairs" (item 3), staying "in a hospital" (item 14), "causing swelling in your ankles or legs" (item 1), and "earn a living" (item 8). However, the magnitude of DIF was determined to be negligible, meaning that no changes to the measure were needed to address DIF. Another DIF evaluation on MLHFQ items was positioned as a "case study" because the research team explored the challenges of evaluating DIF with a small sample size that restricted the team from using standard DIF methods, such as item response theory or logistic regression methods.[40] For both of these studies, sample size was a limitation to drawing cogent conclusions.

A DIF evaluation of the KCCQ has not yet been published, but a study by Hejjaji and colleagues compared psychometric properties of the KCCQ by gender.[41] The results showed that validity, reliability and sensitivity to change are comparable by gender.

Gaps and Future Directions for DIF Analyses in PRO Measures Developed for Adults Diagnosed With Heart Failure

The most conclusive DIF study was the PROMIS+HF study; however, DIF could not be evaluated for all the domains included in the measure due to some domains' being short measures. For the PROMIS+HF domains that could not be quantitatively evaluated, qualitative studies may be a helpful next step to ensure consistency in interpretation of items by subgroups of individuals such as age, gender and race. Questions still remain about potential bias introduced by gender or age DIF on the MLHFQ, and they require confirmation in larger sample sizes to provide more conclusive evidence regarding the magnitude of DIF associated with the 4 items identified by statistically significant DIF. Follow-up qualitative studies may also elucidate whether and how potential interpretation differences by population subgroups may play roles in the MLHFQ. Though the psychometric properties of the KCCQ have been evaluated by gender, a study evaluating DIF analysis has not yet been published for the KCCQ. Similar to PROMIS+HF, the KCCQ includes some short domains that cannot be evaluated for DIF by using quantitative methods. Therefore, further quantitative and qualitative studies could be beneficial for the KCCQ. The MLHFQ is the only measure that has been evaluated with more than 1 DIF method, and more robust DIF analyses are needed.

None of the studies evaluating DIF included DIF analyses specifically for individuals who identify as nonbinary or transgender. Although some cardiovascular research is being conducted in this area,[42,43] it is unclear whether PROMIS+HF might be interpreted differently by individuals who identify as nonbinary or transgender, thus highlighting another gap in current knowledge about health-related quality of life.

In conclusion, at this time, DIF has not conclusively been evaluated for PRO measures used in populations with HF. Without conclusive DIF studies when DIF is suspected, it is unknown whether measurement bias due to DIF influences the ability to validly compare or combine outcomes by subgroups of the population. Conducting formal DIF analyses would be an important step in supporting the interpretation of these instruments as outcomes in trials and as tools for supporting clinical care.

We suggest 2 action items for researchers and clinicians using PRO measures to evaluate outcomes in adults with HF (Fig. 1). The first is to avoid or minimize DIF when developing or revising PRO measures. We can seek to avoid DIF by engaging diverse patient populations in PRO development and validation work, being aware of potential interpretational differences and addressing these differences in early (qualitative) stages. Having a diverse research team and advisors, in terms of race, culture, age, gender, or other characteristics, may provide a strong foundation for identifying DIF hypotheses and addressing them early. The second action item is related to

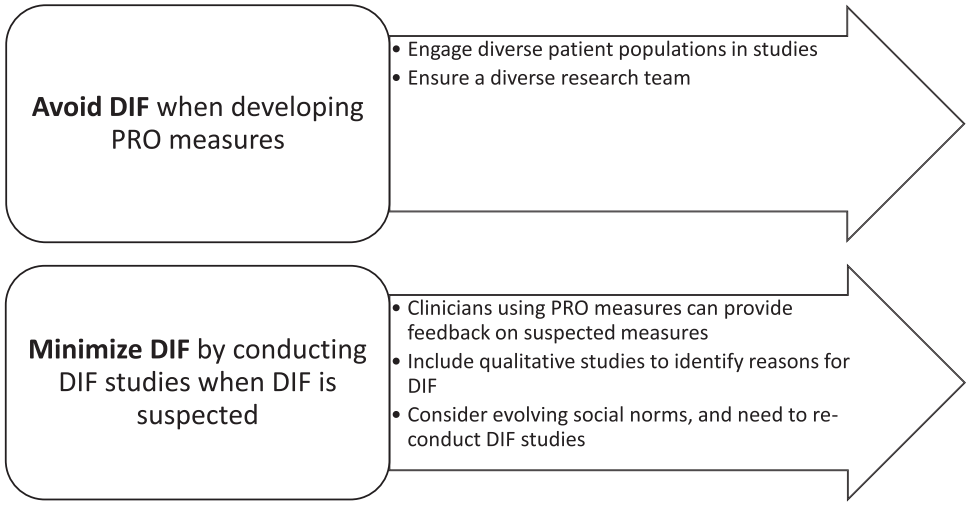

Fig. 1. Action items to avoid or minimize differential item functioning (DIF).

existing PRO measures. We suggest conducting further DIF evaluations for existing PRO measures when sex, race, ethnicity, age, or language (or other population characteristics) when DIF is suspected. Clinicians who administer PRO measures and review PRO scores in clinical care can provide important feedback to developers on potential sources of DIF. Specifically for under-represented and vulnerable groups, qualitative studies may advance understanding of potential interpretational differences or response patterns by various population subgroups, thus refining DIF hypotheses for better understanding of future quantitative results. As social norms evolve, DIF studies may need to be reconducted to capture any differences in interpretation by population subgroups due to the potential evolution of social norms not accounted for in the original development of a PRO measure. An example of such an evolution is the inclusion of individuals who identify as nonbinary or transgender in clinical studies.

A common DIF analysis challenge among existing HF-specific PRO measures is that some PRO measures include domains with 3 or fewer PRO items. This is an advantage in terms of reducing respondent burden and minimizing missing data, but it can limit the evaluation of DIF. Quantitative DIF methods rely on "anchor" items that serve as an estimate of health status for each domain and are assumed to be relatively free of DIF. Without at least 2 viable anchor items, the quantitative methods are unable to be calculated because the estimate of health status (eg, symptoms, physical limitations) is not stable enough for calculating DIF.

Building on the DIF analyses conducted thus far, future studies with larger sample sizes, specific a priori hypotheses and more robust combined qualitative and quantitative steps would provide important evidence to support DIF conclusions.

Conclusion

PROs provide direct information about patients' health status without interpretation from anyone else.[1] In HF, PROs provide important information complementary to clinical measures, patient-defined outcomes,[44] and clinician-reported outcome measures such as New York Heart Association class[45] in the care of patients with HF.[46] It is critical that patients interpret the PRO questions similarly because PRO measures provide unique information in clinical research. DIF occurs when groups of patients (eg, genders, races) respond differently to a particular item of a PRO measure, even after we adjust for their group differences on the outcome being measured. Without formally testing for DIF, it is possible that conclusions drawn from studies using PRO measures are biased, either by revealing subgroup differences that do not actually exist or by masking subgroup differences that do exist. As we strive toward equity in health outcomes, it is important that the tools we use to measure those outcomes are not biased, and DIF evaluations provide a tool for examining bias in PRO measures in HF and beyond. The title of this article is Do PRO Measures Function the Same Way for All Individuals With Heart Failure?, and at this time, the answer is inconclusive. Studies with larger sample sizes, and more robust qualitative steps would further strengthen DIF conclusions for PRO measures developed for use with adults diagnosed with HF.

Disclosures

JS owns the copyright to the KCCQ and receives license fees for the KCCQ. JS receives consulting fees from Bayer, Merck, Novartis, BMS, Janssen, and United Healthcare. JS is supported in grants from Abbott Vascular, Janssen and Myokardia and is on

the Board of Directors for Blue Cross Blue Shield of Kansas City. RM received research support and honoraria from Abbott, American Regent, Amgen, AstraZeneca, Bayer, Boehringer Ingelheim/Eli Lilly, Boston Scientific, Cytokinetics, Fast BioMedical, Gilead, Innolife, Medtronic, Merck, Novartis, Relypsa, Respicardia, Roche, Sanofi, Vifor, and Windtree Therapeutics. TC received research support from Merck and has a consulting agreement with Regenxbio. LL, KW, BR, IP, FB, MT, AS, DH, and BC have no competing interests to disclose.

Funding: This work was supported in part by the U.S. Food & Drug Administration (FDA) Office of Women's Health and the Patient Science & Engagement Program within the FDA's Center for Devices and Radiological Health (Contract 75F40119C10080). This work's contents are solely the responsibility of the authors and do not necessarily represent the official views of the U.S. Department of Health and Human Services (HHS) or FDA. This work was also supported by the Duke University Center for Health Measurement.

Aknowledgments: The authors acknowledge Karen Staman, who provided initial recommendations for medical-editing support.

References

1. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet]. Silver Spring, Md: Food and Drug Administration (US). Updated November 29, 2021. Available at: https://www.ncbi.nlm.nih.gov/books/NBK338448.
2. Heidenreich PA, Fonarow GC, Breathett K, Jurgens CY, Pisani BA, Pozehl BJ, et al. 2020 ACC/AHA Clinical Performance and Quality Measures for Adults With Heart Failure: A Report of the American College of Cardiology/American Heart Association Task Force on Performance Measures. Circ Cardiovasc Qual Outcomes 2020;13:e000099.
3. Anker SD, Agewall S, Borggrefe M, Calvert M, Jaime Caro J, Cowie MR, et al. The importance of patient-reported outcomes: a call for their comprehensive integration in cardiovascular clinical trials. Eur Heart J 2014;35:2001–9.
4. Kelkar AA, Spertus J, Pang P, Pierson RF, Cody RJ, Pina IL, et al. Utility of patient-reported outcome instruments in heart failure. JACC Heart Fail 2016;4:165–75.
5. Green CP, Porter CB, Bresnahan DR, Spertus JA. Development and evaluation of the Kansas City Cardiomyopathy Questionnaire: a new health status measure for heart failure. J Am Coll Cardiol 2000;35:1245–55.
6. Rector TS, Cohn JN. Assessment of patient outcome with the Minnesota Living With Heart-Failure questionnaire: reliability and validity during a randomized, double-blind, placebo-controlled trial of pimobendan. Am Heart J 1992;124:1017–25.
7. Ahmad FS, Kallen MA, Schifferdecker KE, Carluzzo KL, Yount SE, Gelow JM, et al. Development and initial validation of the PROMIS (R)-Plus-HF Profile Measure. Circ Heart Fail 2019;12.
8. US Food and Drug Administration. Medical Device Development Tool (MDDT) qualification decision summary for Minnesota Living with Heart Failure Questionnaire (MLHFQ). 2016; https://www.fda.gov/downloads/MedicalDevices/ScienceandResearch/MedicalDeviceDevelopmentToolsMDDT/UCM581761.pdf.
9. US Food and Drug Administration. Medical Device Development Tool (MDDT) Qualification Decision Summary for Kansas City Cardiomyopathy Questionnaire (KCCQ). Accessed October 13, 2021. 2016.
10. US Food and Drug Administration CDER's COA Qualification Program. DDT COA #000084: Kansas City Cardiomyopathy Questionnaire (KCCQ). 2020 Available at: https://www.fda.gov/drugs/clinical-outcome-assessment-coa-qualification-program/ddt-coa-000084-kansas-city-cardiomyopathy-questionnaire-kccq.
11. McHorney CA, Ware Jr JE, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993;31(3):247–63.
12. Weinfurt KP. Constructing arguments for the interpretation and use of patient-reported outcome measures in research: an application of modern validity theory. Qual Life Res 2021;30(6):1715–22.
13. Teresi JA, Ramirez M, Lai JS, Silver S. Occurrences and sources of differential item functioning (DIF) in patient-reported outcome measures: description of DIF methods, and review of measures of depression, quality of life and general health. Psychol Sci Quart 2008;50:538.
14. Teresi JA, Fleishman JA. Differential item functioning and health assessment. Qual Life Res 2007;16(1):33–42. Suppl.
15. Kitzman DW, Rich MW. Age disparities in heart failure research. JAMA 2010;304:1950–1.
16. Glynn P, Lloyd-Jones DM, Feinstein MJ, Carnethon M, Khan SS. Disparities in cardiovascular mortality related to heart failure in the United States. J Am Coll Cardiol 2019;73:2354–5.
17. Breathett K. Health status equity: a right not a privilege. Am Coll CardiolFoundation, Washington, DC. 2018.
18. Piña IL, Jimenez S, Lewis EF, Morris AA, Onwuanyi A, Tam E, et al. Race and ethnicity in heart failure: JACC focus seminar 8/9. J Am Coll Cardiol 2021;78:2589–98.
19. Shirey TE, Morris AA. Different lenses for the same story: examining how implicit bias can lead us to different clinical decisions for the "same" patient. Am Heart Assoc 2019;:e014355.
20. Khariton Y, Nassif ME, Thomas L, Fonarow GC, Mi X, DeVore AD, et al. Health status disparities by sex, race/ethnicity, and socioeconomic status in outpatients with heart failure. JACC Heart Fail 2018;6:465–73.
21. Daugherty SL, Blair IV, Havranek EP, Furniss A, Dickinson LM, Karimkhani E, et al. Implicit gender bias and the use of cardiovascular tests among cardiologists. J Am Heart Assoc 2017;6:e006872.
22. Dewan P, Rørth R, Jhund PS, Shen L, Raparelli V, Petrie MC, et al. Differential impact of heart failure with reduced ejection fraction on men and women. J Am Coll Cardiol 2019;73:29–40.
23. Adams KF, Dunlap SH, Sueta CA, Clarke SW, Patterson JH, Blauwet MB, et al. Relation between gender, etiology and survival in patients with symptomatic heart failure. J Am Coll Cardiol 1996;28:1781–8.
24. Lee CS, Hiatt SO, Denfeld QE, Chien CV, Mudd JO, Gelow JM. Gender-specific physical symptom biology in heart failure. J Cardiovasc Nurs 2015;30:517–21.

25. McSweeney JC, Cody M, O'Sullivan P, Elberson K, Moser DK, Garvin BJ. Women's early warning symptoms of acute myocardial infarction. Circulation 2003;108:2619–23.
26. Pope JH, Aufderheide TP, Ruthazer R, Woolard RH, Feldman JA, Beshansky JR, et al. Missed diagnoses of acute cardiac ischemia in the emergency department. N Engl J Med 2000;342:1163–70.
27. Heiat A, Gross CP, Krumholz HM. Representation of the elderly, women, and minorities in heart failure clinical trials. Arch Intern Med 2002;162:1682–8.
28. Tahhan AS, Vaduganathan M, Greene SJ, Fonarow GC, Fiuzat M, Jessup M, et al. Enrollment of older patients, women, and racial and ethnic minorities in contemporary heart failure clinical trials: a systematic review. JAMA Cardiol 2018;3:1011–9.
29. Edelen MO, Thissen D, Teresi JA, Kleinman M, Ocepek-Welikson K. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: application to the Mini-Mental State Examination. Med Care 2006;44(11 Suppl 3):S134–42.
30. Crane PK, Gibbons LE, Jolley L, van Belle G. Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. Med Care 2006;44(11 Suppl 3):S115–23.
31. Crane PK, Gibbons LE, Ocepek-Welikson K, Cook K, Cella D, Narasimhalu K, et al. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. Qual Life Res 2007;16(1):69–84. Suppl.
32. Teresi JA. Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. Med Care 2006;44(11 Suppl 3):S152–70.
33. Weinfurt KP, Lin L, Bruner DW, Cyranowski JM, Dombeck CB, Hahn EA, et al. Development and initial validation of the PROMIS® sexual function and satisfaction measures version 2.0. J Sexual Med 2015;12:1961–74.
34. Benitez I, Padilla JL. Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: understanding the causes of differential item functioning by cognitive interviewing. J Mix Meth Res 2014;8:52–68.
35. Reeve BB, Willis G, Shariff-Marco SN, Breen N, Williams DR, Gee GC, et al. Comparing cognitive interviewing and psychometric methods to evaluate a racial/ethnic discrimination scale. Field Meth 2011;23:397–419.
36. Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. Health Qual Life Outcomes 2010;8:81.
37. Edelen MO, Stucky BD, Chandra A. Quantifying "problematic" DIF within an IRT framework: application to a cancer stigma index. Qual Life Res 2015;24:95–103.
38. Zumbo B, Thomas D. A measure of DIF effect size using logistic regression procedures. Philadelphia, PA: Nat Bd Med Examiners, 1996.
39. Munyombwe T, Höfer S, Fitzsimons D, Thompson DR, Lane D, Smith K, et al. An evaluation of the Minnesota living with heart failure questionnaire using Rasch analysis. Qual Life Res 2014;23:1753–65.
40. Sébille V, Auget J-L. Evaluating health-related quality of life: a case-study of differential item functioning analysis in small trials. Communic Statis Theory Meth 2004;33:1403–28.
41. Hejjaji V, Tang Y, Coles T, Jones PG, Reeve BB, Mentz RJ, et al. Psychometric evaluation of the Kansas City Cardiomyopathy Questionnaire in men and women with heart failure. Circulation: Heart Fail 2021;14:e008284.
42. Nota NM, Wiepjes CM, de Blok CJ, Gooren LJ, Kreukels BP, den Heijer M. Occurrence of acute cardiovascular events in transgender individuals receiving hormone therapy: results from a large cohort study. Circulation 2019;139:1461–2.
43. Alzahrani T, Nguyen T, Ryan A, Dwairy A, McCaffrey J, Yunus R, et al. Cardiovascular disease risk factors and myocardial infarction in the transgender population. Circulation: Cardiovasc Qual Outcomes 2019;12:e005597.
44. Sun LY, Rodger J, Duffett L, Tulloch H, Crean AM, Chong AY, et al. Derivation of patient-defined adverse cardiovascular and noncardiovascular events through a modified delphi process. JAMA Netw Open 2021;4:e2032095.
45. Heidenreich PA. The growing case for routine collection of patient-reported outcomes. JAMA Cardiol 2021;6:497–8.
46. Coles TM, Hernandez AF, Reeve BB, Cook K, Edwards MC, Boutin M, et al. Enabling patient-reported outcome measures in clinical trials, exemplified by cardiovascular trials. Health Qual Life Outcomes 2021;19:1–7.