Women's Health Issues 29-S1 (2019) S12–S14

WOMEN'S HEALTH ISSUES
www.whijournal.com

Commentary

Designing Studies for Sex and Gender Analyses: How Research Can Derive Clinically Useful Knowledge for Women's Health

Ruth Klap, PhD [a,b,*], Keith Humphreys, PhD [c]

[a] *VA HSR&D Center for the Study of Healthcare Innovation, Implementation and Policy, VA Greater Los Angeles Health Care System, Los Angeles, California*
[b] *Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California*
[c] *Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California*

Article history: Received 13 May 2019; Accepted 14 May 2019

It was once considered acceptable for clinical researchers to draw conclusions about women's health based on studies that only enrolled men (Wizeman, 2012). Fortunately, in recent decades, government, research funders, and other groups have promoted greater inclusion of women participants in clinical research (Freeman et al., 2017; Society for Women's Health Research, 2001). As a result, the participation of women in clinical research has increased (Freeman et al., 2017; Lippman, 2006; Society for Women's Health Research, 2001), and the articles in this special issue of *Women's Health Issues* are some of the fruits of that progress. Yet many studies within and outside Veterans Affairs (VA) do not report results by sex or gender[1] (Avery & Clark, 2016; Clayton & Tannenbaum, 2016; Duan-Porter et al., 2016; Vidaver, Lafleur, Tong, Bradshaw, & Marts, 2000). Some scholars propose increasing the frequency of such analyses by having journals or funders mandate them (Mazure & Jones, 2015; Sugimoto, Ahn, Smith, Macaluso, & Larivière, 2019).

In this commentary, we express reservations about that approach and instead suggest that more clinically relevant lessons about women's health could be derived from research if 1) studies are adequately designed and powered for testing sex and/or gender interactions in advance, 2) study exclusion criteria are chosen in light of their sex and gender impact, 3) researchers use research designs and statistical techniques that can illuminate sex and gender differences, and 4) researchers routinely make data publicly available after publication.

Mandatory Analysis of Clinical Research Outcomes by Sex and Gender: Some Reservations

Clinical trials estimate average treatment effects (Kravitz, Duan, & Braslow, 2004), and ample evidence highlights the hazards of assuming such effects apply equally to male and female participants (Bailey, 2007; Hawkes, Haseen, & Aounallah-Skhiri, 2019; Wallach et al., 2017). Yet we doubt the usefulness of a policy of funding agencies or journals mandating or endorsing analysis by sex and gender, for all studies, for several reasons.

Only a minority of studies are designed and conducted in a manner that adequately supports the study of outcome differences among sample subgroups (Bucholz & Krumholz, 2015; Freeman et al., 2017), in part because no benchmark defining adequate enrollment by sex has been established by the U.S. Food and Drug Administration or the National Institutes of Health (Bucholz & Krumholz, 2015). Most clinical trials are only powered to identify overall treatment effects. A trial that has 80% power to detect an overall treatment effect will only have 29% power to detect an interaction effect of the same magnitude; sample size would have to increase four-fold for the study to have 80% power to detect a sex- or gender-specific effect (Hernández, Boersma, Murray, Habbema, & Steyerberg, 2006).

Dr. Klap's effort was funded by VA Health Services Research and Development Service (HSR&D) Women's Health Research Network (Project # SDR 10-012). Dr. Humphreys' effort was funded by a Senior Research Career Scientist Award (RCS 04-141) from the VA HSR&D.

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect official positions of the Department of Veterans Affairs. The funding source had no involvement in any aspects of the study design, data collection, analysis or interpretation; writing of the article; or decision to submit the article for publication. The authors have no financial conflicts of interest to report.

* Correspondence to: Ruth Klap, PhD, VA HSR&D Center for the Study of Healthcare Innovation, Implementation and Policy, VA Greater Los Angeles Healthcare System, 11301 Wilshire Blvd (151) Building 206, Los Angeles, CA 90073. Phone: (310) 478-3711x44636; fax: (818) 895-5838.

E-mail address: Ruth.Klap@va.gov (R. Klap).

[1] The term "sex" is used to refer to biological or physiological factors and "gender" is used to refer to psychosocial, gender identity, or cultural factors. It should be noted that these terms are not mutually exclusive (Clayton & Tannenbaum, 2016).

1049-3867/$ - see front matter Published by Elsevier Inc. on behalf of Jacobs Institute of Women's Health.
https://doi.org/10.1016/j.whi.2019.05.002

False negatives are always a risk when conducting underpowered analyses; as the number of statistical comparisons conducted within a data set increases, the likelihood of finding statistically significant but clinically meaningless differences increases. Richard Peto (2011) puckishly illustrated the perils of running multiple unplanned comparisons by demonstrating that although daily aspirin was clearly beneficial overall, it was harmful for people born under Libra and Gemini astrological signs.

Journal editors or research funders mandating reanalysis of data to account for sex and gender will not address the problems associated with the design and conduct of a study (Wizeman, 2012). Few studies provide a rationale or a hypothesis related to the subgroup analyses (Aulakh & Anand, 2007; Avery & Clark, 2016). Evidence suggests that most published subgroup analyses are misleading (Aulakh & Anand, 2007; Avery & Clark, 2016; Petticrew et al., 2012; Wallach et al., 2017). Attempts at replicating subgroup findings are rare, but when done, the original findings are not generally supported (Wallach et al., 2017), and many subgroup analyses noted in published abstracts are not supported by their own data (Wallach et al., 2017). Researchers often incorrectly conclude that a statistically significant effect in one subgroup but not another (stratified analyses) establishes evidence of a subgroup affect when, in fact, such analyses require the conduct of a formal interaction test (Wallach et al., 2017). Furthermore, subgroup analyses of one variable at a time often fail to detect treatment effect differences because patients have many attributes that can influence treatment effectiveness (Kent & Hayward, 2007). In light of these issues, we are wary of requiring sex and gender analysis as a general matter in clinical research or gauging progress in the inclusion of women in research through the existence of reported subgroup analyses. We share the concern of Springer, Stellman, and Jordan-Young (2012) that this practice could lead to an ever-growing catalogue of differences (that are likely spurious) that will distract from needed research into specific mechanisms that lead to sex or gender differences and the relationship between sex and gender differences. Furthermore, such requirements are at odds with statistical and methodological guidance regarding planning for subgroup analyses.

Designing Studies for Sex and Gender Subgroup Analyses

Planning for subgroup analysis is infinitely superior to implementing unplanned post hoc analyses (Aulakh & Anand, 2007; Yusuf, Wittes, Probstfield, & Tyroler, 1991). Proper planning requires having an explicit rationale for performing subgroup analyses, specifying sex- and gender-related hypotheses a priori, ensuring sufficient statistical power, stratifying randomizations by sex, designing a formal test of the interaction, and adjusting for multiple statistical tests.

Achieving a sample size with sufficient power for subgroup analyses is resource intensive and thus requires a strong rationale (Wizeman, 2012). There are costs and benefits associated with powering studies for sex and gender subgroup analyses in the general population, and these tradeoffs are exacerbated in the VA where women represent only 7.5% of the patient population (Frayne et al., 2018). Although the VA Women's Health Practice-Based Research Network can increase the representation of women veterans in VA research (Frayne et al., 2013), even with a multisite approach it can be challenging to sufficiently power studies of VA users for sex and gender comparisons, depending on the research question and study design.

It is therefore critical to identify research questions where sex or gender is likely to matter and adequately design and power studies for subgroup analyses in these situations. "A starting assumption that there is a sex difference for any association being studied could lead to publication of false conclusions," Wizeman (2012) cautions. Researchers need to consider what is known about sex and gender differences at the research design stage. Information on how groups based on sex and age metabolize medications, for instance, could allow for the identification of questions where subgroup differences are possible and should be studied and those where they are unlikely to occur (Allmark, 2004).

Choosing Exclusion Criteria Wisely in Light of Disparate Sex and Gender Effects

Clinical researchers often undermine the clinical relevance of their data by choosing trial exclusion criteria that disproportionately prevent certain subsets of women from enrolling in studies. The most obvious example is a demographic one, namely excluding the elderly (who are disproportionately women), a practice of 40% of clinical trials in medicine (Zulman et al., 2011). Exclusion of participants with disorders more common in women (e.g., depression; Humphreys & Williams, 2018), or social conditions more common in women (e.g., unemployment; Humphreys, Weingardt, & Harris, 2007) can have the same effect.

Importantly, although oversampling women can improve power for subgroup analyses, differences in the sex and gender impact of an exclusion criterion cannot be compensated for by oversampling. This method only produces a larger relatively unrepresentative sample of women, allowing more precision in estimating the wrong answer. For example, a study that excluded individuals under 5′7″ tall would have a more representative sample of men than women, no matter whether women were oversampled or not.

Research Designs and Statistical Techniques that Can Illuminate Sex and Gender Differences

Because the conduct of well-planned, sufficiently powered subgroup analyses is often not feasible, alternative approaches are needed. Although it is risky to act on findings from exploratory subgroup analyses, based on interaction tests, they can be used, if interpreted cautiously, to generate hypotheses that are tested in future studies (see Brown et al., [2019] and Danan et al., [2019]). Furthermore, the risks of producing type 1 errors (false positives) can be mitigated by correcting for multiple comparisons (see Benjamini & Hochberg [1995] and Naylor et al., [2019]). Although stratified analyses are not good indicators of whether sex or gender differences exist, if all studies routinely presented findings stratified by sex or gender, perhaps as supplementary materials with warnings about overinterpretation of findings, power issues could be addressed through meta-analysis techniques (Wizeman, 2012). Faced with selective reporting by gender, however, meta-analyses will be seriously compromised by publication bias (Bailey, 2007).

Suitable ways of addressing sex and gender subgroup differences include using underutilized statistical techniques (Wizeman, 2012). Bayesian approaches, for instance, can be used to reduce sample size requirements in traditional frequentist analyses (Berry, 2005). In addition, Bayesian and adaptive analytic trial designs can be used to update trial information as results accumulate, allowing for a reduction in sample size requirements and a focus on clinically relevant subgroups (Berry, 2005; Kravitz et al., 2004). The use of risk stratification is another

promising alternative to subgroup analyses. Specifically in situations where an externally developed risk prediction tool is available, multivariate risk models often have greater power to detect treatment differences than individual subgroup analyses (Hayward, Kent, Vijan, & Hofer, 2006; Kent & Hayward, 2007; Wizeman, 2012). By combining multiple patient attributes into a single score that describes a single dimension of risk upon which a treatment effect is likely to vary, risk-stratified analyses minimize difficulties associated with multiple comparisons and poor statistical power (Kent & Hayward, 2007).

Data Sharing Can Help

If researchers make raw data available, then postpublication data could be pooled across multiple studies that address the same condition. Data merging could be especially valuable if common measures were included in studies (Wizeman, 2012). Researchers could increase statistical power by pooling data across studies that individually enrolled small numbers of women veterans.

Conclusion

The importance of gathering high-quality data on women's responses to health care interventions cannot be overstated from a scientific, clinical, or ethical viewpoint. However, mandating post hoc sex and gender analysis may just as easily result in misleading rather than useful data on women's health. A better path forward involves a priori considerations of sex and gender analysis where warranted, choosing exclusion criteria wisely, exploiting innovative statistical techniques, and broadly sharing data so that other researchers can analyze large pools of data on women's response to treatments.

References

Allmark, P. (2004). Should research samples reflect the diversity of the population? *Journal of Medical Ethics*, 30(2), 185–189.

Aulakh, A. K., & Anand, S. S. (2007). Sex and gender subgroup analyses of randomized trials: The need to proceed with caution. *Women's Health Issues*, 17(6), 342–350.

Avery, E., & Clark, J. (2016). Sex-related reporting in randomised controlled trials in medical journals. *Lancet*, 388(10062), 2839–2840.

Bailey, K. R. (2007). Reporting of sex-specific results: A statistician's perspective. *Mayo Clinic Proceedings*, 82(2), 158.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300.

Berry, D. A. (2005). *Introduction to Bayesian methods III: Use and interpretation of Bayesian tools in design and analysis*. Thousand Oaks, CA: Sage.

Brown, M. C., Sims, K. J., Gifford, E., Goldstein, K. M., Johnson, M. R., Williams, C., & Provenzale, D. (2019). Gender-based differences among 1990-91 Gulf War Era veterans: Demographics, lifestyle behaviors, and health conditions. *Womens Health Issues*, 29(Suppl.1), S47–S55.

Bucholz, E. M., & Krumholz, H. M. (2015). Women in clinical research: What we need for progress. *Circulation. Cardiovasc Quality and Outcomes*, 8(Suppl. 2), S1–S3.

Clayton, J. A., & Tannenbaum, C. (2016). Reporting sex, gender, or both in clinical research? *JAMA*, 316(18), 1863–1864.

Danan, E. R., Sherman, S. E., Clothier, B., Burgess, D. J., Pinsker, E., Joseph, A. M., … Fu, S. F. (2019). Smoking cessation among female and male veterans before and after a randomized trial of proactive outreach. *Womens Health Issues*, 29(Suppl.1), S15–S23.

Duan-Porter, W., Goldstein, K. M., McDuffie, J. R., Hughes, J. M., Clowse, M. E., Klap, R. S., … Gierisch, J. M. (2016). Reporting of sex effects by systematic reviews on interventions for depression, diabetes, and chronic pain. *Annals of Internal Medicine*, 165(3), 184–193.

Frayne, S. M., Carney, D. V., Bastian, L., Bean-Mayberry, B., Sadler, A., Klap, R., … Yee, E. F. (2013). The VA women's health practice-based research network: Amplifying women veterans' voices in VA research. *Journal of General Internal Medicine*, 28(2), 504–509.

Frayne, S. M., Phibbs, C., Saechao, F., Friedman, S., Shaw, J., Romodan, Y., … Haskel, S. (2018). Sourcebook: Women Veterans in the Veterans Health Administration. In. *Volume 4: Longitudinal Trends in Sociodemographics, Utilization, Health Profile, and Geographic Distribution. Women's Health Evaluation Initiative*. Washington, DC: Women's Health Services, Veterans Health Administration, Department of Veterans Affairs.

Freeman, A., Stanko, P., Berkowitz, L. N., Parnell, N., Zuppe, A., Bale, T. L., … Epperson, C. N. (2017). Inclusion of sex and gender in biomedical research: Survey of clinical research proposed at the University of Pennsylvania. *Biology of Sex Differences*, 8(1), 22.

Hawkes, S., Haseen, F., & Aounallah-Skhiri, H. (2019). Measurement and meaning: Reporting sex in health research. *Lancet*, 393(10171), 497–499.

Hayward, R. A., Kent, D. M., Vijan, S., & Hofer, T. P. (2006). Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Medical Research Methodology*, 6(1), 18.

Hernández, A. V., Boersma, E., Murray, G. D., Habbema, J. D. F., & Steyerberg, E. W. (2006). Subgroup analyses in therapeutic cardiovascular clinical trials: Are most of them misleading? *American Heart Journal*, 151(2), 257–264.

Humphreys, K., Weingardt, K. R., & Harris, A. H. (2007). Influence of subject eligibility criteria on compliance with National Institutes of Health guidelines for inclusion of women, minorities, and children in treatment research. *Alcoholism: Clinical and Experimental Research*, 31(6), 988–995.

Humphreys, K., & Williams, L. M. (2018). What can treatment research offer general practice? *Lancet Psychiatry*, 5(4), 295–297.

Kent, D. M., & Hayward, R. A. (2007). Limitations of applying summary results of clinical trials to individual patients: The need for risk stratification. *JAMA*, 298(10), 1209–1212.

Kravitz, R. L., Duan, N., & Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Quarterly*, 82(4), 661–687.

Lippman, A. (2006). *The inclusion of women in clinical trials: Are we asking the right questions?* Toronto, Canada: Women and Health Protection.

Mazure, C. M., & Jones, D. P. (2015). Twenty years and still counting: Including women as participants and studying sex and gender in biomedical research. *BMC Womens Health*, 15(1), 94.

Naylor, J. C., Wagner, H. R., Johnston, C., Elbogen, E. E., Brancu, M., Marx, C. E., … Strauss, J. L. (2019). Pain intensity and pain interference in male and female Iraq/Afghanistan-era veterans. *Women' Health Issues*, 29(Suppl.1), S24–S31.

Peto, R. (2011). Current misconception 3: That subgroup-specific trial mortality results often provide a good basis for individualising patient care. *British Journal of Cancer*, 104(7), 1057–1058.

Petticrew, M., Tugwell, P., Kristjansson, E., Oliver, S., Ueffing, E., & Welch, V. (2012). Damned if you do, damned if you don't: Subgroup analysis and equity. *Journal of Epidemiology and Community Health*, 66(1), 95–98.

Society for Women's Health Research. (2001). Institute of medicine report validates the science of sex differences. *Journal of Womens Health and Gender Based Medicine*, 10(4), 303–304.

Springer, K. W., Stellman, J. M., & Jordan-Young, R. M. (2012). Beyond a catalogue of differences: A theoretical frame and good practice guidelines for researching sex/gender in human health. *Social Science & Medicine*, 74(11), 1817–1824.

Sugimoto, C. R., Ahn, Y.-Y., Smith, E., Macaluso, B., & Larivière, V. (2019). Factors affecting sex-related reporting in medical research: A cross-disciplinary bibliometric analysis. *Lancet*, 393(10171), 550–559.

Vidaver, R. M., Lafleur, B., Tong, C., Bradshaw, R., & Marts, S. A. (2000). Women subjects in NIH-funded clinical research literature: Lack of progress in both representation and analysis by sex. *Journal of Womens Health and Gender Based Medicine*, 9(5), 495–504.

Wallach, J. D., Sullivan, P. G., Trepanowski, J. F., Sainani, K. L., Steyerberg, E. W., & Ioannidis, J. P. (2017). Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. *JAMA Internal Medicine*, 177(4), 554–560.

Wizeman, T. M. (2012). *Sex-specific reporting of scientific research: A workshop summary*. Washington, DC: National Academies Press.

Yusuf, S., Wittes, J., Probstfield, J., & Tyroler, H. A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*, 266(1), 93–98.

Zulman, D. M., Sussman, J. B., Chen, X., Cigolle, C. T., Blaum, C. S., & Hayward, R. A. (2011). Examining the evidence: A systematic review of the inclusion and analysis of older adults in randomized controlled trials. *Journal of General Internal Medicine*, 26(7), 783–790.

Author Descriptions

Ruth Klap, PhD, is a Research Health Scientist, VA HSR&D Center for the Study of Healthcare Innovation, Implementation and Policy, and National Consortium Program Manager, VA Women's Health Research Network and an Associate Research Sociologist at the UCLA David Geffen School of Medicine.

Keith Humphreys, PhD, is a Senior Research Career Scientist at the VA HSR&D Center on Innovation to Implementation, VA Palo Alto Health Care System, and is the Esther Ting Memorial Professor in the Department of Psychiatry and Behavioral Sciences at Stanford University.