PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cohort Profile: The Women's Health Accelerometry Collaboration
AUTHORS	Evenson, Kelly; Bellettiere, John; Cuthbertson, Carmen; Di, Chongzhi; Dushkes, Rimma; Howard, Annie; Parada, Humberto; Schumacher, Benjamin; Shiroma, Eric J.; Wang, Guangxing; Lee, I-Min; LaCroix, Andrea

VERSION 1 – REVIEW

REVIEWER	Huisingh-Scheetz, Megan University of Chicago
REVIEW RETURNED	08-Jun-2021

GENERAL COMMENTS	Acknowledgements: Dr. Nabiel Mir, a fellow-in-training, contributed to this review. Overview: The purpose of this paper is to present a merged cohort profile of two cohort studies in women and to guide potential data users on how to best align measures for studying cancer-related outcomes. It meets criteria for publication as a Cohort Profile analysis. I overall think the paper presents a unique cohort as it represents one of the few (or only) datasets that include highly granular cancer incidence information with objective physical activity monitoring. The data are limited to women. Very few chronic diseases and other comorbid conditions or cancer-related risk factors are presented in this cohort profile, unclear if available?. Specific suggestions about how the authors might improve the value of this paper for potential data users are included below. Abstract: Would be helpful to have the available cancer incidence rate(s) in the abstract, the primary planned outcome. Intro: P5 line 50. Authors report a moderate association b/w physical activity and lung cancer then in the next sentence report a limited dose response gradient b/w physical activity and lung cancer. Please clarify. P5 line 52-56. Please clarify the following points: 1) The review indicated limited evidence on activity occurring outside of leisure-time…". Are the authors saying that leisure time activity WAS significantly associated with cancer risk but moderate-vigorous activity was not? 2) "….lack of associations w/ physical activity by subgroups…". Are the authors saying that the significant association between free-living activity and cancer risk cited in line 43 was no longer significant when looking at subgroups? Or that there was no evidence for differential effects across subgroups? The goal statement is very clear which is to target late life cancer incidence and to identify major risk factors such as physical activity. This is clinically meaningful sample as it is a group bearing

	the burden of cancer incidence and yet this group undergoes the least amount of cancer screening. Cohort description: Missing setting and location of both cohorts (required by BMJ open). Follow up period is also not described BMJ Open recommends a flow diagram to record how many people were approached/retained. Would consider adding. P9 line 3. WHS: can the authors include information on what proportion of those eventually enrolled in the accelerometry study were also in the prior pharmacotherapy intervention arm for preventing CV disease and cancer? How might this prior pharmacotherapy exposure impact any potential associations between accelerometry and cancer? Also, can the authors provide more detail about where WHS study participants were recruited from? P9 line 41. "From this sample…." The text reads as though "sample" refers to the group who consented to the WHI/OPACH ancillary study. However, the sentence content ("58 died before recruitment….") suggests the authors are describing a pre-consented group (eg, how can people die before recruitment but after consenting? How can people be consented but ineligible?). Please clarify. Also, can the authors more detail about where WHI/OPACH study participants were recruited from? P11 line 20. Can the authors provide a citation supporting their definition of a moderate-vigorous bout of activity? Allowing up to 5 consecutive minutes in a 10-minute window below the moderate activity count cut-point seems like a flexible definition. P13 line 24. Please provide question details or references for the following self-reported measures: general health, average walking speed. P13 line 29. Did participants report any other chronic conditions besides DM and CAD? How about other cancer-related behaviors and exposures like diet, sexual behaviors, drug use or environmental toxins? If so, please include. The descriptions of the accelerometry and cancer data collection are extremely useful for potential data users and well described. Given this is an older adult study focusing on physical activity and cancer incidence, are there other meaningful covariates on important geriatric syndromes that you can include that would impact activity? For example, did the studies collect any objective measures of physical function (usual walk, TUG, SPPB), disability or cognition? P 12 line 41. The authors do not include basic analytic information for how the descriptive data for this paper were generated. Specifically, were survey weights used / needed or other cohort-specific analytic considerations for potential data users? Findings to date: Would consider providing descriptive differences at baseline for a few select accelerometry parameters between lung vs breast cancer vs non cancer (tabulated or mentioned in text). P15 line 41. The authors note that "some of the differences between cohorts may be due to age or other confounders…..". Age could easily be explored in this paper by reporting sedentary behavior time by age categories. If the age-adjusted values are aligned, it would support merging data from both studies. Table 3. Would add % weekend vs week day wear (eg, report 0, 1 or 2 weekend days included in wear time) as activity is significantly different by this factor. The authors provide helpful detail on the number of participants in each study who agreed to participate in the accelerometry

	substudies as well as those who dropped out / didn't wear the device. Can the authors also report how those people who contributed accelerometry data were different from those who did not? It is likely that those who participated are healthier than those who did not. This information will be relevant to future data users and would be frequently referenced. Strengths and Limitations: It would help overall if the discussion here could be framed to better support future data users. For example, the authors might provide some analytic considerations for modeling in light of the cohort differences. Another potential example: would the authors recommend data users include a study-specific indicator variable in their models? Other ideas: they could make suggestions on potentially important covariates to include in models relating activity to cancer in this dataset. They might comment on how the cohorts differ (or not) from the general US population of older women. Line 29-30 last paragraph p17. Would provide reference data from Table 3 (and refer readers to Table 3) for this statement: "Despite these differences, there did not appear to be differential impact on accelerometer awake wear time between the cohorts"

REVIEWER	Borghese, Michael Health Canada
REVIEW RETURNED	12-Jun-2021

GENERAL COMMENTS	Journal: BMJ Open Manuscript ID: bmjopen-2021-052038 Cohort Profile: The Women's Health Accelerometry Collaboration Reviewed: June 11th, 2021 Overview: In this paper the authors describe the development of the Women's Health Accelerometry Collaboration, a harmonization of two prospective cohort studies. I have recommended a major revision of the paper for the reasons outlined below. The paper would benefit from presenting a rationale for the need for a harmonized cohort earlier on in the introduction. For example, the following statement from the strengths/limitations section could be moved up: "The Women's Health Accelerometry Collaboration cohort's primary strength is the statistical power to be able to address research questions regarding physical activity and sedentary behaviour among older women in a cost-efficient manner by using data from existing studies." In the spirit of this "primary" strength, I would have expected to see one or more power calculations to support this claim. Since 17,061 of the 23,443 women come from a single cohort, could the authors have not just used data from this cohort? The addition of 6382 participants, from the perspective of statistical power, seems somewhat trivial in this context. I suggest that the authors identify one to three primary outcomes that they plan on studying and perform power calculations for these outcomes under expected disease incidence scenarios – specifically, does a sample size of 23,443 participants provide adequate power, while a sample size of 17,061 does not? This collaboration would benefit from having follow-up accelerometer data. A single measure of activity at baseline is

	rather limited when predicting diseases with a long latency period, especially with age-related changes in activity levels over this interval. Repeated measures, throughout the at-risk period (i.e., over decades in some cases), would be tremendously beneficial. The authors mention repeated data collection over a 2-3 year period, and it is good to see that these are reproducible over the short-term, but it's not clear to me why they wouldn't plan for future data collection. I am concerned that these samples may be too different to combine. There are some notable differences in sociodemographic, lifestyle, and chronic disease prevalence at baseline. However, there are drastic differences in physical activity levels between the samples – a nearly 2-fold difference in the mean. Moreover, the levels of MVPA in these samples of older women are almost unbelievable, especially in the WHS cohort (mean 91.9 min/day of MVPA). This could be because the physical activity decision rules/cut-points were developed using the WHI/OPACH cohort only in an internal calibration study. Applying these cut-points to the WHS cohort and observing such a drastic (and possibly unbelievable) difference could mean that: 1) the cut-points/method is too specific to WHI/OPACH cohort and is not stable in other cohorts, or 2) the WHI/OPACH and WHS cohorts are too different to combine - specifically, that they may be representative of different source populations. To resolve this, I suggest that the authors consider another, externally derived, set of cut-points to help rule out the possibility of the 2nd option. Alternatively, can the authors provide evidence of these cut-points being used in another sample to support their stability across samples? Can the authors please explain how each of these older cohorts can obtain such high levels of average min/day of MVPA alongside such low average daily step counts? Their MVPA is well above nearly any physical guideline in the world, even pediatric ones for the WHS cohort, and yet the step count is less than or around half of the 10000 step/day guideline. The authors may want to revisit their cut-points and decision rules (perhaps the short epoch length is the issue?). Time spent sedentary and in light physical activity seem to be typical, so it could just be the moderate cut-point that's the issue. Specific comments/suggestions: Strengths and limitations of this study Suggest removing the reference to men for the 4th limitation: "This cohort is limited to women age 62 years and older, and thus may not be generalizable to men or to younger adults". This is not a limitation of the harmonized cohort since the source population consists of only women by definition. Cohort description For a cohort profile paper, I think it's important to describe in detail the characteristics of those who agreed to participate vs. those who declined vs. did not respond for each of the cohorts. Specifically, did those who declined have a higher/lower prevalence of cancer diagnosis at baseline? It would be good to know about any differences in sociodemographic (e.g., ethnicity, income, education), lifestyle (e.g., smoking, alcohol use) or chronic conditions (e.g., diabetes, hypertension, etc.). The concern is that only the healthiest or most privileged women might agree to

	participate in such a study. This could be done in Table 2 by adding a column beside the "missing" category, or as a separate table. Accelerometer data collection "For women with missing sleep log data, their in-bed and out-of-bed times were imputed using person-specific means, if available, or the sample mean." The within-person mean is not great but it could be the best that's available, but imputing the sample mean is definitely not a good idea. I think this should be re-visited. There are a number of sleep onset/offset algorithms for waist-worn ActiGraphs that are available, although mostly in children. Have the authors considering applying such algorithms to identify sleep time in cases where the log data are not available? 15s epochs seem rather short for this population. This is what is typically used in children whereas 1 minute epochs are more common in adults. The Choi algorithms used in this paper were developed with 1 minute epochs. Why did the authors select 15s epochs? The validation paper by the same author (Evenson) used 15s epochs, but the rationale provided from the Pettee Gabriel et al. (2010) paper is not well supported, since they showed that 60s epochs resulted in stronger associations between MVPA and a number of health metrics. "The bout must start and end with moderate-to-vigorous physical activity." I'm wondering if the authors could elaborate on this. Of course it must start with MVPA, but I'm not sure I understand how all bouts can be forced to end with MVPA? Or does this mean that the "tails" of the bouts are trimmed off if they're below threshold? It might be worth briefly clarifying this in the paper. Cancer Incidence and Mortality Outcomes This section is outside of the scope of my expertise. It seems appropriately detailed to me, but it should be reviewed by someone with expertise in cancer diagnosis. I would like to know how this collaboration compares with existing cohorts that are capable of examining physical activity/cancer associations. A paragraph dedicated to contextualing the novelty of this cohort would be beneficial. Other measurements "Height and weight were self-reported in WHS and measured in WHI/OPACH." This is a limitation worth noting I think. How was walking speed measured? Was this directly measured or reported? Findings to date Accelerometer adherence appears to be rather good. This could be emphasized more. I disagree that the proportion of coronary heart disease between the studies is similar (10.1% WHI/OPACH, 4.3% WHS). This is a >2-fold difference. I suggest revising.

	I think there are some notable differences between these samples that should be identified as limitations for harmonization – notably the use of the postmenopausal hormones and diabetes. Table 1 Suggest adding % retained at each stage alongside the n Table 2 Suggest comparing the two studies statistically as well (e.g., Chi-square, t-test). This doesn't have to be complicated, but it will allow the reader identify potentially notable differences/similarities.

VERSION 1 – AUTHOR RESPONSE

Response to Reviewer Comments: Manuscript #bmjopen-2021-052038

"Cohort Profile: The Women's Health Accelerometry Collaboration"
We thank the reviewers for their helpful comments that served to improve the paper. Please see our responses to each comment next. Changes to the text were indicated by burgundy font color.

Reviewer 1:
Overview: The purpose of this paper is to present a merged cohort profile of two cohort studies in women and to guide potential data users on how to best align measures for studying cancer-related outcomes. It meets criteria for publication as a Cohort Profile analysis. I overall think the paper presents a unique cohort as it represents one of the few (or only) datasets that include highly granular cancer incidence information with objective physical activity monitoring. The data are limited to women. Very few chronic diseases and other comorbid conditions or cancer-related risk factors are presented in this cohort profile, unclear if available?. Specific suggestions about how the authors might improve the value of this paper for potential data users are included below.

1. Abstract: Would be helpful to have the available cancer incidence rate(s) in the abstract, the primary planned outcome.
Response: We added the following sentence to the abstract: During an average of 5.9 (SD 1.6) years of follow-up, 1378 cancer events among which 414 were fatal have occurred.

2. Intro: P5 line 50. Authors report a moderate association b/w physical activity and lung cancer then in the next sentence report a limited dose response gradient b/w physical activity and lung cancer. Please clarify.
Response: The Physical Activity Guidelines Committee separately scored the overall association from whether or not there was a graded dose response association found. The information is summarized in table 1 from reference #5. We have edited the sentences in the paper to try to better clarify. Specifically, lung cancer was rated as moderate overall evidence grade and limited dose response grade.

3. P5 line 52-56. Please clarify the following points: 1) The review indicated limited evidence on activity occurring outside of leisure-time…". Are the authors saying that leisure time activity WAS significantly associated with cancer risk but moderate-vigorous activity was not?
Response: This sentence reflected on the type of activity and not intensity. We have clarified the sentence to now state: The review indicated limited evidence on physical activity occurring outside of leisure-time, such as transportation, occupational, or household activities.

4. “….lack of associations w/ physical activity by subgroups…”. Are the authors saying that the significant association between free-living activity and cancer risk cited in line 43 was no longer significant when looking at subgroups? Or that there was no evidence for differential effects across subgroups?
Response: Few studies reported any exploration into subgroups. We have clarified the sentence to now state: The review also found that few studies reported on associations between physical activity and cancer by population subgroups, such as by age, socioeconomic status, or race/ethnicity.

5. The goal statement is very clear which is to target late life cancer incidence and to identify major risk factors such as physical activity. This is clinically meaningful sample as it is a group bearing the burden of cancer incidence and yet this group undergoes the least amount of cancer screening.
Response: Thank you.

6. Cohort description: Missing setting and location of both cohorts (required by BMJ open). Follow up period is also not described BMJ Open recommends a flow diagram to record how many people were approached/retained. Would consider adding.
Response: The setting and location are now present in both the abstract and paper. The follow-up period is described in the methods under the section “Cancer Incidence and Mortality Outcomes”. The actual follow-up period is provided in the “Findings to Date” section and has now additionally been added to the abstract. Since the flow diagram is optional based on BMJ Open guidelines, and since we were asked to include percentages to the table, we have opted to retain the table format.

7. P9 line 3. WHS: can the authors include information on what proportion of those eventually enrolled in the accelerometry study were also in the prior pharmacotherapy intervention arm for preventing CV disease and cancer? How might this prior pharmacotherapy exposure impact any potential associations between accelerometry and cancer? Also, can the authors provide more detail about where WHS study participants were recruited from?
Response: All of the women in the WHS accelerometry study were previously in the pharmacotherapy intervention arms (either active or placebo). We previously reported no significant associations of the interventions on cancer incidence and mortality (JAMA 2005;294:47-55 and 56-65); thus, the interventions are unlikely to affect the accelerometry-cancer associations we seek to investigate. This information is now added to the paper. WHS participants were recruited from throughout the United States. This was added (see comment #6).

8. P9 line 41. “From this sample….” The text reads as though “sample” refers to the group who consented to the WHI/OPACH ancillary study. However, the sentence content (“58 died before recruitment….”) suggests the authors are describing a pre-consented group (eg, how can people die before recruitment but after consenting? How can people be consented but ineligible?). Please clarify. Also, can the authors more detail about where WHI/OPACH study participants were recruited from?
Response: “From this sample” refers to the 8618 women who consented to be contacted for the OPACH study and 58 died before consenting to WHI/OPACH. We have revised the phrasing. The WHI study initially recruited women 50-79 years for either a clinical trial(s) or an observational study from 40 clinical sites throughout the United States from 1993 to 1998. This text was also added to the paper.

9. P11 line 20. Can the authors provide a citation supporting their definition of a moderate-vigorous bout of activity? Allowing up to 5 consecutive minutes in a 10-minute window below the moderate activity count cut-point seems like a flexible definition.
Response: We originally stated that “a moderate-to-vigorous bout was defined as >=10 minutes of consecutive moderate-to-vigorous minutes with allowance for interruptions for up to 20% of the time below the threshold and <5 consecutive minutes below the threshold. The bout must start and end with moderate-to-vigorous physical activity.” The designation for “<5 consecutive minutes below the

threshold" is applied as a constraint for longer bouts (i.e., >=25 minutes), based on the length and percent of bout time above the threshold. As an example, for a 30 minute MVPA bout, 20% that can fall below the cutpoint would equal 6 minutes. However, we capped it at 5 minutes so as not to allow lengthy periods below. This definition came from extensive work examining where and how MVPA bouts occurred from a sample of adults that wore an accelerometer and GPS concurrently for 3 weeks. We added the references (Holliday et al, 2017a; Holliday et al, 2017b) and tried to further clarify the bout definition in the text.

10. P13 line 24. Please provide question details or references for the following self-reported measures: general health, average walking speed.
Response: We have provided details on the measurement of general health and walking speed to the text.

11. P13 line 29. Did participants report any other chronic conditions besides DM and CAD? How about other cancer-related behaviors and exposures like diet, sexual behaviors, drug use or environmental toxins? If so, please include.
Response: We have included all measures that were similarly ascertained for both cohorts near the time of accelerometry measurement and able to be harmonized. For example, diet was assessed using different instruments at different time periods and therefore was not included.

12. The descriptions of the accelerometry and cancer data collection are extremely useful for potential data users and well described. Given this is an older adult study focusing on physical activity and cancer incidence, are there other meaningful covariates on important geriatric syndromes that you can include that would impact activity? For example, did the studies collect any objective measures of physical function (usual walk, TUG, SPPB), disability or cognition?
Response: This is a good question that we considered extensively before combining the cohorts. The WHI/OPACH study collected measures of physical function close to the time of accelerometry measurement during the in-person interview. However, WHS relied on self-report and did not collect similar measures close to the time of accelerometry measurement.

13. P 12 line 41. The authors do not include basic analytic information for how the descriptive data for this paper were generated. Specifically, were survey weights used / needed or other cohort-specific analytic considerations for potential data users?
Response: Both cohorts were women who volunteered to participate; they do not represent an underlying population. Thus, survey weights were not generated.

14. Findings to date: Would consider providing descriptive differences at baseline for a few select accelerometry parameters between lung vs breast cancer vs non cancer (tabulated or mentioned in text).
Response: The aims of the paper were to describe the methods and process of harmonizing data across the cohorts in order to explore this question. We feel it is beyond the scope of the paper to provide information on associations between physical activity and cancer type, a topic we are working on for subsequent manuscripts.

15. P15 line 41. The authors note that "some of the differences between cohorts may be due to age or other confounders…..". Age could easily be explored in this paper by reporting sedentary behavior time by age categories. If the age-adjusted values are aligned, it would support merging data from both studies.
Response: Thank you for this suggestion. Table 4 was added to provide accelerometry measures by tertile of age for both cohorts.

16. Table 3. Would add % weekend vs week day wear (eg, report 0, 1 or 2 weekend days included in wear time) as activity is significantly different by this factor.
Response: This information was added to Table 3. A few WHS women wore the accelerometer for more than 7 days, which spanned over more than 2 weekend days.

17. The authors provide helpful detail on the number of participants in each study who agreed to participate in the accelerometry substudies as well as those who dropped out / didn't wear the device. Can the authors also report how those people who contributed accelerometry data were different from those who did not? It is likely that those who participated are healthier than those who did not. This information will be relevant to future data users and would be frequently referenced.

Response: This information was previously reported for WHI/OPACH comparing those who return the accelerometer with data vs. those who did not (LaCroix et al., 2017) and comparing those participants in and not in the study (Jain et al., 2020 – see Supplementary Table 1). For example, women who were included in the analysis were younger, reported higher general health, had fewer number of chronic conditions, and higher physical function score compared to women not included in the study. Similarly for WHS, participants included were younger and generally healthier, with fewer chronic conditions.

18. Strengths and Limitations: It would help overall if the discussion here could be framed to better support future data users. For example, the authors might provide some analytic considerations for modeling in light of the cohort differences. Another potential example: would the authors recommend data users include a study-specific indicator variable in their models? Other ideas: they could make suggestions on potentially important covariates to include in models relating activity to cancer in this dataset. They might comment on how the cohorts differ (or not) from the general US population of older women.
Response: The information on modeling is provided in the "Statistical Analysis" section of the paper. We added a sentence to clarify that the confounding could be assessed from the list of covariates described.

19. Line 29-30 last paragraph p17. Would provide reference data from Table 3 (and refer readers to Table 3) for this statement: "Despite these differences, there did not appear to be differential impact on accelerometer awake wear time between the cohorts"
Response: Changed as requested.

Reviewer 2:
Overview: In this paper the authors describe the development of the Women's Health Accelerometry Collaboration, a harmonization of two prospective cohort studies. I have recommended a major revision
of the paper for the reasons outlined below.

1. The paper would benefit from presenting a rationale for the need for a harmonized cohort earlier on in the introduction. For example, the following statement from the strengths/limitations section could be moved up: "The Women's Health Accelerometry Collaboration cohort's primary strength is the statistical power to be able to address research questions regarding physical activity and sedentary behaviour among older women in a cost-efficient manner by using data from existing studies."
Response: We did not move the statement up to the introduction of the paper, only because it would be redundant with the required section on "strengths and limitations".

2. In the spirit of this “primary” strength, I would have expected to see one or more power calculations to support this claim. Since 17,061 of the 23,443 women come from a single cohort, could the authors have not just used data from this cohort? The addition of 6382 participants, from the perspective of statistical power, seems somewhat trivial in this context. I suggest that the authors identify one to three primary outcomes that they plan on studying and perform power calculations for these outcomes under expected disease incidence scenarios – specifically, does a sample size of 23,443 participants provide adequate power, while a sample size of 17,061 does not?
Response: While WHS does have a larger sample size, statistical power in the case of survival analysis is dependent not only on the sample size of individuals at risk but also on the number of events. Moreover, over the next few years the statistical power should continue to increase with longer follow-up time in which more events occur. While statistical power would be, as the reviewer suggests, relatively similar for high incidence cancers such as breast cancer, for other cancers subtypes that are less common, adding in the additional 6,489 individuals corresponds to a total of 754 incidence cancer cases. This can substantially improve statistical power. For example, at this point in time, for lung cancer if we used only data from WHS we would have a total of 98 incident cases, but if we combine data from WHS and WHI/OPACH we have a total of 146

incident cases. By adding in these additional changes, we can now detect hazard ratios for lung cancer as small as 1.26 for a 1 standard deviation change in physical activity as compared to hazard ratios as small as 1.35 if we excluded these individuals. While studies investigating the associations between these more rare cancer subtypes and physical activity have been limited due to small sample sizes and few events, this improvement in statistical power, allows us to be better equipped to investigate these associations in this population of older women. In addition to increasing power for these less common cancer outcomes, by including both cohorts we are capturing more diversity in the population of women in this age range and therefore this allows us to better understand these associations in a more heterogenous population. These strengths were clarified in the “Strengths” section of the paper.

3. This collaboration would benefit from having follow-up accelerometer data. A single measure of activity at baseline is rather limited when predicting diseases with a long latency period, especially with age-related changes in activity levels over this interval. Repeated measures, throughout the at-risk period (i.e., over decades in some cases), would be tremendously beneficial. The authors mention repeated data collection over a 2-3 year period, and it is good to see that these are reproducible over the short-term, but it’s not clear to me why they wouldn’t plan for future data collection.
Response: We agree that follow-up accelerometer data would enhance the study. However, this requires further funding and time. In the interim, we are optimistic that our study can provide important insight into the proposed associations.

4. I am concerned that these samples may be too different to combine. There are some notable differences in sociodemographic, lifestyle, and chronic disease prevalence at baseline. However, there are drastic differences in physical activity levels between the samples – a nearly 2-fold difference in the mean. Moreover, the levels of MVPA in these samples of older women are almost unbelievable, especially in the WHS cohort (mean 91.9 min/day of MVPA). This could be because the physical activity decision rules/cut-points were developed using the WHI/OPACH cohort only in an internal calibration study. Applying these cut-points to the WHS cohort and observing such a drastic (and possibly unbelievable) difference could mean that: 1) the cut-points/method is too specific to WHI/OPACH cohort and is not stable in other cohorts, or 2) the WHI/OPACH and WHS cohorts are too different to combine - specifically, that they may be representative of different source populations. To resolve this, I suggest that the authors consider another, externally derived, set of cut-points to help rule out the possibility of the 2nd option. Alternatively, can the authors provide evidence of these cut-points being used in another sample to support their stability across samples?

Response: We do not believe that the cutpoints developed are specific only to WHI/OPACH women. In fact, they remain the most applicable cutpoints to use among women of this age group with this accelerometer (ActiGraph GT3X) and wear position (hip). The calibration study included 200 women, but not all of them were WHI participants.

5. Can the authors please explain how each of these older cohorts can obtain such high levels of average min/day of MVPA alongside such low average daily step counts? Their MVPA is well above nearly any physical guideline in the world, even pediatric ones for the WHS cohort, and yet the step count is less than or around half of the 10000 step/day guideline. The authors may want to revisit their cut-points and decision rules (perhaps the short epoch length is the issue?). Time spent sedentary and in light physical activity seem to be typical, so it could just be the moderate cut-point that's the issue.
Response: The cutpoint is calibrated to estimate moderate to vigorous activity among older women, which is why the number of minutes is higher than reported on other studies that utilize calibration equations developed in younger samples of adults (i.e., what might be a "light" activity in a younger woman may actually require moderate or higher effort in an older woman). We have added this clarification to the strengths section of the paper.

6. Strengths and limitations of this study: Suggest removing the reference to men for the 4th limitation: "This cohort is limited to women age 62 years and older, and thus may not be generalizable to men or to younger adults". This is not a limitation of the harmonized cohort since the source population consists of only women by definition.
Response: This bullet was removed as requested.

7. Cohort description: For a cohort profile paper, I think it's important to describe in detail the characteristics of those who agreed to participate vs. those who declined vs. did not respond for each of the cohorts. Specifically, did those who declined have a higher/lower prevalence of cancer diagnosis at baseline? It would be good to know about any differences in sociodemographic (e.g., ethnicity, income, education), lifestyle (e.g., smoking, alcohol use) or chronic conditions (e.g., diabetes, hypertension, etc.). The concern is that only the healthiest or most privileged women might agree to participate in such a study. This could be done in Table 2 by adding a column beside the "missing" category, or as a separate table.
Response: This information was previously reported for WHI/OPACH comparing those who return the accelerometer with data vs. those who did not (LaCroix et al., 2017) and comparing those participants in and not in the study (Jain et al., 2020 – see Supplementary Table 1). For example, women who were included in the analysis were younger, reported higher general health, had fewer number of chronic conditions, and higher physical function score compared to women included in the study. Similarly for WHS, participants included were younger and generally healthier, with fewer chronic conditions.

8. Accelerometer data collection: "For women with missing sleep log data, their in-bed and out-of-bed times were imputed using person-specific means, if available, or the sample mean." The within-person mean is not great but it could be the best that's available, but imputing the sample mean is definitely not a good idea. I think this should be re-visited. There are a number of sleep onset/offset algorithms for waist-worn ActiGraphs that are available, although mostly in children. Have the authors considering applying such algorithms to identify sleep time in cases where the log data are not available? 15s epochs seem rather short for this population. This is what is typically used in children whereas 1 minute epochs are more common in adults.
Response: Thank you for this comment. We conducted a laboratory-based study among 200 older women that was specifically designed to calibrate our accelerometers to measure physical activity and sedentary behavior among older women using 15-second epoch data. Regarding automated sleep algorithms, our team evaluated existing algorithms, which at the time were all developed among

children or young adults, and adapted the algorithms to perform better among older adults (Bellettiere, J., Y. Zhang, V. Berardi, K. M. Full, J. Kerr, M. J. LaMonte, K. R. Evenson, M. Hovell, A. Z. LaCroix and C. Di (2019). "Parameterizing and validating existing algorithms for identifying out-of-bed time using hip-worn accelerometer data from older women." Physiol Meas 40(7): 075008). After parameterizing the existing algorithms, there was good agreement between the algorithm-identified in-bed times and the self-reported in-bed times, however, there was a wide 95% range of agreement even with the best performing algorithm, which was up to +- 5 hour per day. In a subsequent study evaluating the association of sedentary behavior and cardiovascular disease, we published a sensitivity analysis where sleep time was imputed using the population mean vs. using the best performing automated algorithm that we specifically tuned for our older women (Bellettiere, J., M. J. LaMonte, K. R. Evenson, E. Rillamas-Sun, J. Kerr, I. M. Lee, C. Di, D. E. Rosenberg, M. Stefanick, D. M. Buchner, M. F. Hovell and A. Z. LaCroix (2019). "Sedentary behavior and cardiovascular disease in older women: The Objective Physical Activity and Cardiovascular Health (OPACH) Study." Circulation 139(8): 1036-1046.). The resulting sedentary behavior and cardiovascular disease association was virtually the same regardless of the imputation method used. Therefore, in our nearly 20 published OPACH studies, we have used the mean imputation method to address missing sleep log data, which is present for just ~5% of the OPACH sample. Because 1) algorithms are costly to apply and are still error-prone even when tuned to the target population, 2) imputation using the sample mean and population-tuned algorithms yielded similar results in a previous study, and 3) to be consistent with our previous studies, we feel that for the small number of people with missing sleep logs (~5%), using the sample mean is our preferred imputation method.

9. The Choi algorithms used in this paper were developed with 1 minute epochs. Why did the authors select 15s epochs? The validation paper by the same author (Evenson) used 15s epochs, but the rationale provided from the Pettee Gabriel et al. (2010) paper is not well supported, since they showed that 60s epochs resulted in stronger associations between MVPA and a number of health metrics.
Response: We selected 15-seconds to enhance accuracy and to align with the cutpoints developed from the calibration study conducted on women of the same age. When applying the Choi algorithms, we did aggregate the data into 1 minute epochs and then apply the algorithms, for the reason that the reviewer pointed out.

10. “The bout must start and end with moderate-to-vigorous physical activity.” I’m wondering if the authors
could elaborate on this. Of course it must start with MVPA, but I’m not sure I understand how all bouts can be forced to end with MVPA? Or does this mean that the “tails” of the bouts are trimmed off if they’re below threshold? It might be worth briefly clarifying this in the paper.
Response: The MVPA bout must end with MVPA, so as not to erroneously count physical activity below moderate intensity or sedentary behavior. Because we allow for some proportion of time to drop below the threshold (to mimic, for example, waiting at a stoplight while taking a walk), we do not count the time that drops below the threshold at the end of the bout. We added two references on the MVPA bout definition and tried to further clarify.

11. Cancer Incidence and Mortality Outcomes: This section is outside of the scope of my expertise. It seems appropriately detailed to me, but it should be reviewed by someone with expertise in cancer diagnosis. I would like to know how this collaboration compares with existing cohorts that are capable of examining physical activity/cancer associations. A paragraph dedicated to contextualing the novelty of this cohort would be beneficial.
Response: Only a few cohort studies can address accelerometer-assessed physical activity and sedentary behavior with cancer outcomes among older women. The novelty of the cohort is addressed in the “Strengths” section of the paper.

12. Other measurements: “Height and weight were self-reported in WHS and measured in WHI/OPACH.” This is a limitation worth noting I think.
Response: We have added this as a limitation to the paper.

13. How was walking speed measured? Was this directly measured or reported?
Response: Walking speed was self-reported by both cohorts. We added a detailed paragraph about how the questions were harmonized.

14. Findings to date: Accelerometer adherence appears to be rather good. This could be emphasized more.
Response: We added the excellent adherence to the “Strengths” section of the paper.

15. I disagree that the proportion of coronary heart disease between the studies is similar (10.1% WHI/OPACH, 4.3% WHS). This is a >2-fold difference. I suggest revising.
Response: We now note the difference in coronary heart disease in the text of the results.

16. I think there are some notable differences between these samples that should be identified as limitations for harmonization – notably the use of the postmenopausal hormones and diabetes.

Response: While there are differences between cohorts, these differences bring more diversity to the research. Moreover, we will consider all potential covariates as confounders and can adjust for these differences statistically. We will also treat cohort as a stratification variable in the model, and will explore results separately for the two cohorts.

17. Table 1: Suggest adding % retained at each stage alongside the n
Response: Percents were added as requested to Table 1.

18. Table 2: Suggest comparing the two studies statistically as well (e.g., Chi-square, t-test). This doesn’t have to be complicated, but it will allow the reader identify potentially notable differences/similarities.
Response: Due to the sample size, a very small difference will be statistically different. Nevertheless, as requested we did calculate p values and provide a note at the bottom of Table 2, reflecting that every association between WHS and WHI/OPACH had a $p<0.0001$.

VERSION 2 – REVIEW

REVIEWER	Borghese, Michael Health Canada
REVIEW RETURNED	07-Sep-2021

GENERAL COMMENTS	Thanks for the detailed responses. All the best.