PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Rationale and design of a cohort Study On Primary ovarian insufficiency in female survivors of Hodgkin lymphoma: Influence on long-term Adverse effects (SOPHIA)
AUTHORS	Krul, Inge; Opstal-van Winden, Annemieke; Zijlstra, Josée; Appelman, Yolande; Schagen, SB; Meijboom, Lilian; Serné, Erik; Lambalk, Cornelis; Lips, Paul; van Dulmen - den Broeder, Eline; Hauptmann, M; Daniëls, Laurien; Aleman, Berthe; van Leeuwen, Flora

VERSION 1 – REVIEW

REVIEWER	Giovanni Codacci-Pisanelli University of Rome "la Sapienza", Italy
REVIEW RETURNED	01-Aug-2017

GENERAL COMMENTS	I would suggest analysing patients that received pelvic radiotherapy as a separate group.

REVIEWER	Charles L. Shapiro MD, FASCO Icahn School of Medicine at Mt Sinai, New York, New York. United States
REVIEW RETURNED	08-Oct-2017

GENERAL COMMENTS	This protocol is designed for female Hodgkin’s disease (HD) survivors at least 8+ years after receiving their initial treatment. The aims are to assess the long-term and late effects, including bone density, cognitive function, cardiovascular disease, and the use hormone replacement after menopause, in those women who or did not experience primary ovarian failure (POF). There is a separate ongoing study of women without HD but with POF that will serve as controls for cardiovascular disease. Abstract: page 3, line 27: It is not clear how the criteria “treated since 1965 and followed for 8 or more years” were chosen. Abstract: page 3, line 28: The “women will be invited" likely introduces selection bias. To overcome this problem, the authors should record the same demographic, treatment, and other on-study variables in women who decline to participate. The point is are these "selected survivors" representative of the overall group of survivors. Abstract: page 4, line 1: The results MAY help, instead of will help. Page 5, lines 19-35: Although age at the time of treatment may not influence the rate of POF, the age of patients receive treatment may.

Although the authors’ state in their previous study the median age of POF was 33 years, they fail to provide a range and the median age and range of the population is not provided. The study is not likely powered to answer a question related to the age at treatment. Thus, a 29-year-old may be very different than a 39 year with POF, regarding the main outcome variables.

Pages 5-6, lines 59-2: There are a paucity studies in HD survivors and bone density. However, there are many studies in premenopausal women with chemotherapy-induced POF and bone loss. The diseases and treatment are different, but why should the acute estrogen deprivation of POF and bone loss differ between these populations? These may impact the number of tests they propose for bone mineral density described in Table 1. If authors feel that POF and bone loss is different between Hodgkin’s and breast cancer survivors, then they should explain their rationale.

Page 7, line 15: The references don't support the statement that "...POF many years after treatment." Reference 21 is a study of oophorectomy patients and the rate of bone. Reference 51 is a 1996 review article of women with breast cancer summarizing the data on amenorrhea after chemotherapy. Some amenorrhea is permanent, and some women recover menstrual function after a period amenorrhea. POF is closer to having oophorectomy than natural menopause, which occurs for some years. POF doesn't happen many years after initial treatment.

Page 7, line 41-60, the paragraph under Aims: It would be clearer to state the hypotheses one by one. The hypothesis of reverse causality is intriguing, however, to consider as a primary hypothesis is unfounded by the lack of data. It should a secondary or exploratory hypothesis. As mentioned above, the power not likely to be adequate to separate the "acute versus gradual POF groups." It can set as an exploratory hypothesis, but then at a minimum, the definitions of acute and gradual need to be defined.

The above leads a broader question. Is POF defined by patient recall, or do you have actual data on your definitions of menopause listed on page 8 lines 59-2 on page 9, which includes ≥ four months of amenorrhea with two serum FSH levels in postmenopausal range or ≥ twelve months of amenorrhea before the age 40? Which of the former or latter do you plan to rely on is not clear. If using the former, you have the problem of recall bias, and this impacts profoundly on the primary aim of assessing the late and long-term effects in women who do and did not develop POF.

Pages 10, 11, Table 1: Asking patients to remember their last menstrual period and menopausal age, or medical documentation of these dates would be fraught with problems if the proposed study were performed in the US. Perhaps, in Netherlands it is different. If so, the protocol should state what is found in the medical records concerning menstrual function and age of menopause

If the design of the study was a prospective cohort design, perhaps you could make a case for the many biomarkers. But in observational cross-sectional design, many of the biomarkers are either not useful in answering the primary hypotheses or the data is conflicting on the individual biomarker. For example, a one-time assessment of bone turnover markers, CRP, NT-pro BNP is

	unnecessary. Also, the EKG and ECHO are not needed if you plan on performing the CCTA and IMT. Although the neurocognitive measures are justified, the problem is you don't know what the baseline pre-treatment neurocognitive function was. For example, studies that suggest neurocognitive in breast cancer patients pre-chemotherapy is abnormal in some domains. Page 13, lines 1-15, the paragraph under medical records: Changing treatment practices, which authors acknowledge, could have a significant effect on outcome variables, but it is not likely the study is adequately powered to account the de-escalation of treatment that has occurred over time. Page 14, lines 1-13, the paragraph under blood samples. Banking blood samples are great idea. Running these biomarkers in one-time assessment is not. Page 16, lines 2: The analysis of T-scores, not bone density, is the important variable. Everyone loses bone, but the T-scores are predictive fractures and is a more meaningful endpoint. Page 16, lines 31-60, the paragraph under statistical analyses: Should the p-values be more rigorous (lower) because of the potential multiple comparisons problem? Or is the sample size large enough to perform all of the proposed analyses? In summary, this study proposal strikes me as too ambitious with a multitude of questions on a relatively small, selected group of survivors. How is POF defined is a critical issue and the results are subject to recall bias or lack of documentation. If this not the case, then the authors should state the accuracy of the data sources determining POF. The protocol includes unnecessary testing, and the failure to separate the primary hypotheses from secondary exploratory ones. Also, this protocol would benefit from an independent statistical review to ensure they have an adequate patient population to answer the primary hypotheses.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:
Reviewer Name
Giovanni Codacci-Pisanelli

Institution and Country
University of Rome "la Sapienza"
Italy

Please state any competing interests or state ‘None declared’:
none declared

1. I would suggest analyzing patients that received pelvic radiotherapy as a separate group.

Reply: Thank you for your comment. We are indeed interested in the effect of a RT-induced primary ovarian insufficiency (POI) on the main outcomes and we will therefore perform a subgroup analysis. In addition, we will evaluate the difference between a CT- and RT-induced POI in more detail. This was already stated in the original manuscript in the section Statistical analyses.

Reviewer: 2

Reviewer Name
Charles L. Shapiro MD, FASCO

Institution and Country
Icahn School of Medicine at Mt Sinai, New York, New York. United States

Please state any competing interests or state 'None declared':
None declared

This protocol is designed for female Hodgkin's disease (HD) survivors at least 8+ years after receiving their initial treatment. The aims are to assess the long-term and late effects, including bone density, cognitive function, cardiovascular disease, and the use hormone replacement after menopause, in those women who or did not experience primary ovarian failure (POF). There is a separate ongoing study of women without HD but with POF that will serve as controls for cardiovascular disease.

1. Abstract: page 3, line 27: It is not clear how the criteria "treated since 1965 and followed for 8 or more years" were chosen.

Reply: The first criterion, treated since 1965, is based on the definition of our study cohort as participants will be recruited from a large previously described cohort of HL survivors treated in the Netherlands between 1965 and 2000. Registry data on HL patients treated before 1965 are not available.
However, we deliberately chose to include women who survived 8 years or more while the cohort consists of 5-year HL survivors, because we are interested in the long-term effects of POI. We do not expect a short-term effect of POI on our main outcomes.

We added two sentences to our Methods section 'Design and study population' to clarify these criteria.

Added sentence 1: Registry data on HL patients treated before 1965 are not available.

Added sentence 2: The latter criterion was chosen because we are interested in the long-term effects of POI.

2. Abstract: page 3, line 28: The "women will be invited" likely introduces selection bias. To overcome this problem, the authors should record the same demographic, treatment, and other on- study variables in women who decline to participate. The point is are these "selected survivors" representative of the overall group of survivors.

Reply: We agree with the reviewer that it is important to check whether the included women are representative of the overall group of 8-year HL survivors. Since this study is nested within a large cohort, we already have data available for all 8-year HL survivors on general patient characteristics, HL treatment and some follow-up data (vital status, diagnosis of second malignancies and cardiovascular disease). This enables us to examine possible differences in patients who are included and those who decline. Also, we will be able to examine whether eligible 8-year survivors who died before study invitation died due to one of our outcomes of interest. Due to the high risk of late adverse effects in HL survivors, some women are already deceased

If it would turn out that a relatively large proportion of patients in the POI group (compared to the comparison group) has died of CVD, we will be able to report this, which is a big advantage in a cross-sectional study.

Of course, for recruitment we are dependent on the women who will visit the Survivorship Care Clinics as women will be invited during their visit. Some women are under surveillance in a local hospital rather than their original HL treatment centre, while women who are (feeling) healthy may not visit the Survivorship Care Clinic because of medical costs and/or other obligations (e.g., work, family). This is inherent to research in clinical practice, but we will put efforts into obtaining medical data for women under surveillance in others hospitals. Moreover, the Survivorship Care Clinics actively recall women with different HL treatment regimens.

We already mentioned this important point in the original manuscript in the fourth and fifth paragraph of the Discussion section. We further clarified this point by adjusting the Methods section in the abstract and by adding/adjusting two sentences in the Methods section ‘Design and study population’.

Adjusted sentence in Abstract: Women visiting a survivorship care outpatient clinic (SSC) will be invited for a neurocognitive, cardiovascular and BMD assessment, and asked to complete several questionnaires and to provide a blood sample.

Adjusted sentence in Methods section 1: The SOPHIA-study is an observational cross-sectional study among female HL survivors who are being followed in an outpatient survivorship care clinic. Participants will be invited for a neurocognitive, cardiovascular and BMD assessment, and asked to complete several questionnaires and to provide a blood sample.

Added sentence in Methods section 2: General patient characteristics, HL treatment data and follow-up data on vital status, second malignancies and CVD are already available for all 8-year HL-survivors in these three hospitals, enabling us to monitor possible differences between patients who participate and those who decline. Also, we will be able to examine whether eligible 8-year survivors who died before study invitation died due to one of our outcomes of interest. Due to the high risk of late adverse effects in HL survivors, some women are already deceased. If it would turn out that a relatively large proportion of patients in the POI group (compared to the comparison group) has died of CVD, we will be able to report this, which is a big advantage in a cross-sectional study.

3. Abstract: page 4, line 1: The results MAY help, instead of will help.

Reply: Thank you for your comment. We have adjusted the sentence.

Old sentence: Results of this study will help to identify and timely refer those Hodgkin lymphoma survivors who are at increased risk for osteoporosis, neurocognitive dysfunction and possibly cardiovascular disease due to treatment-induced primary ovarian insufficiency for interventions in order to reduce morbidity and enhance quality of life.

New sentence: Results of this study may help to identify and timely refer those Hodgkin lymphoma survivors who are at increased risk for osteoporosis, neurocognitive dysfunction and possibly cardiovascular disease due to treatment-induced primary ovarian insufficiency for interventions in order to reduce morbidity and enhance quality of life.

4. Page 5, lines 19-35: Although age at the time of treatment may not influence the rate of POF, the age of patients receive treatment may. Although the authors' state in their previous study the median age of POF was 33 years, they fail to provide a range and the median age and range of the population is not provided. The study is not likely powered to answer a question related to the age at treatment. Thus, a 29-year-old may be very different than a 39 year with POF, regarding the main outcome variables.

Reply: We do not entirely understand this comment of the reviewer. It seems that some words have been left out. We agree with the reviewer that both age at HL treatment and age at developing POI may influence the main outcomes, and our study is powered to examine the independent effect of both age variables. We will do this by running Cox regression models with age as a time scale. It is likely, however, that we do not have enough power to look at the modifying effect of age at POI. This is not our main aim of the study and was therefore not described.

We added a sentence to the Section Statistical analyses:

Added sentence: Cox regression models with age as a time scale will be used to examine the independent effect of age at HL treatment and age at developing POI.

The range of age at POI in the mentioned study was 19-39 years. We added this range to the sentence in the Background section.

Adjusted sentence: In our earlier study among female 5-year HL survivors treated between 1965 and 1995, women developed POI at a median age of 33 years (range 19-39 years).

5. Pages 5-6, lines 59-2: There are a paucity studies in HD survivors and bone density. However, there are many studies in premenopausal women with chemotherapy-induced POF and bone loss. The diseases and treatment are different, but why should the acute estrogen deprivation of POF and bone loss differ between these populations? These may impact the number of tests they propose for bone mineral density described in Table 1. If authors feel that POF and bone loss is different between Hodgkin's and breast cancer survivors, then they should explain their rationale.

Reply: This is an important question. We have considered whether any medical test could be left out, as participants in this study are asked to undergo quite some medical tests. However, a large part of the described tests are part of routine medical care, as shown in detail in figure 1.
For women with POI, a DEXA-scan is part of routine medical care, so the DEXA-scan would only be additional for women without POI in our comparison group. We are aware that multiple studies already have examined the association between POI and bone mineral density in breast cancer survivors and in women with a natural early menopause. However, the large majority of HL survivors are diagnosed at a young age and consequently develop POI at a younger age than breast cancer survivors or women with a natural POI. Hence, the extent of reduced bone mineral density may be different, both with medium-term and long-term follow-up.
In addition, we have the possibility to combine the DEXA-scan with an instant vertebral assessment to gain insight into vertebral fracture risk. So far, not many studies have reported on this outcome. Since fracture risk is related to bone mineral density, we feel it would be incomplete to not report on bone mineral density as well.
Finally, we feel that the DEXA-scan poses a very small burden on our study population.

We adjusted the paragraph in the Background section to emphasize the need to examine bone mineral density and fracture risk specifically in HL survivors.

Old paragraph: So far, many studies have been conducted in women with an early natural or surgery-induced menopause, while studies evaluating the long-term effects of CT- and/or RT-induced POI on BMD and fracture risk are limited. Two small studies among HL survivors reported a significantly reduced BMD after treatment-induced POI [31 32], while another study found no association.[33]. More research is needed to identify the extent of reduced BMD and prevalence of osteoporotic fractures among female HL survivors who developed POI.

New paragraph: So far, many studies have been conducted among breast cancer survivors or in women with an early natural or surgery-induced menopause, while studies evaluating the long-term effects of CT- and/or RT-induced POI on BMD and fracture risk in HL survivors are limited. Two small studies among HL survivors reported a significantly reduced BMD after treatment-induced POI [31 32], while another study found no association.[33] Since HL survivors develop POI at a younger age than breast cancer survivors or the general population, more research is needed to identify the extent of reduced BMD and prevalence of osteoporotic fractures among female HL survivors who developed POI.

6. Page 7, line 15: The references don't support the statement that "...POF many years after treatment." Reference 21 is a study of oophorectomy patients and the rate of bone. Reference 51 is a 1996 review article of women with breast cancer summarizing the data on amenorrhea after chemotherapy. Some amenorrhea is permanent, and some women recover menstrual function after a period amenorrhea. POF is closer to having oophorectomy than natural menopause, which occurs for some years. POF doesn't happen many years after initial treatment.

Reply: In our manuscript, we refer to references 20 and 51 after our statement: Moreover, the majority of studies looked at the effects after oophorectomy, characterized by an abrupt drop of oestrogen levels, while oestrogen levels may decrease gradually in CT-induced POI occurring many years after treatment [20 51]. This does not include reference 21, which is indeed about bone mineral density.

Reference 20 (Hendrix et al 2005), reports the following: “Bilateral oophorectomy and premature ovarian failure are similar in that they are both hypo-estrogenic conditions that increase a woman’s risk for osteoporotic fracture and coronary heart disease (CHD). However, they are different in that with premature ovarian failure, the menopausal transition is similar to natural menopause; there is a gradual decline in sex hormone levels, and women often experience intermittent menopausal symptoms and irregular uterine bleeding for many years before the onset of amenorrhea. In contrast, surgical castration results in acute hypo-estrogenism and hypo-androgenism, most commonly resulting in acute symptomatology. Estrogen requirements for symptom control in women with bilateral oophorectomy may be higher than in the woman with premature ovarian failure, and the inability to discontinue estrogen also may be greater in the woman who has undergone oophorectomy because of the severe symptomatology.”

Reference 50 (Bines et al, 1996) states the following: “Few investigators have considered the possibility that chemotherapy may cause sublethal ovarian damage, resulting in premature menopause months or years after completion of treatment”.

The remark that POI can occur many years after initial treatment is supported by our own preliminary data on POI in HL survivors. The median duration of ovarian function after HL treatment in women who developed POI is 4 years, with an interquartile range of 1-10 years.
Therefore, it would be very interesting to look at a possible difference between an acute and gradually developed POI. We adjusted the sentence in the Background section.

Adjusted sentence: Moreover, the majority of studies looked at the effects after oophorectomy, characterized by an abrupt drop of oestrogen levels, while oestrogen levels may decrease gradually in CT-induced POI occurring many years after treatment [20 51]. Preliminary data on POI in HL survivors within our cohort show that women who developed POI had a median duration of ovarian function after HL treatment of 4 years (interquartile range 1-10 years).

7. Page 7, line 41-60, the paragraph under Aims: It would be clearer to state the hypotheses one by one. The hypothesis of reverse causality is intriguing, however, to consider as a primary hypothesis is unfounded by the lack of data. It should a secondary or exploratory hypothesis. As mentioned above, the power not likely to be adequate to separate the "acute versus gradual POF groups." It can set as an exploratory hypothesis, but then at a minimum, the definitions of acute and gradual need to be defined.

Reply: To make the distinction between our primary aim and our secondary aims more clear, we have adjusted this section by stating our primary aim followed by our secondary aims one by one.

A requirement of BMJ open is to report a clear statement of the main study aim and major hypothesis/research question. Our primary aim is to examine the long-term effects of treatment-induced POI on BMD, cardiovascular status, neurocognitive function and QoL.
Since cardiovascular status is our primary outcome, we wanted to formulate a clear hypothesis. Based on the existing literature regarding the association between POI and CVD risk among women in the general population, the conventional view is that reproductive hormone deprivation after POI increases CVD risk. However, there are two interesting alternative hypotheses (1. Risk factors for CVD determine menopausal age and 2. Accelerated biological aging underlies both early menopause and increased CVD risk), and we think that our study population is one of the rare populations where these intriguing hypotheses can be tested. We therefore hypothesize that CVD risk is not increased in female HL survivors with POI compared to HL survivors without POI. Based on our sample size calculations, we should have sufficient power to address this.

With regard to the definition of acute and gradual, we define acute POI as development of POI within 1 year after HL treatment. A gradually developed POI occurs 1 year or later after treatment. Indeed, this is a secondary aim and involves an exploratory hypothesis, as we are not certain that we will have sufficient power to address this issue. We added this definition to the secondary aims section and the note that we may not have enough power.

Old sentence: To examine whether long-term effects differ between acute and more gradually developed POI.

New sentence: To examine whether long-term effects differ between acute (<1 year after HL treatment) and more gradually (≥ 1 year after HL treatment) developed POI if there is sufficient power.

8. The above leads a broader question. Is POF defined by patient recall, or do you have actual data on your definitions of menopause listed on page 8 lines 59-2 on page 9, which includes ≥ four months of amenorrhea with two serum FSH levels in postmenopausal range or ≥ twelve months of amenorrhea before the age 40? Which of the former or latter do you plan to rely on is not clear. If using the former, you have the problem of recall bias, and this impacts profoundly on the primary aim of assessing the late and long-term effects in women who do and did not develop POF.

Reply: The women who will be invited to participate in this study received HL treatment in a broad range of calendar years: 1965-2000. Therefore, we will make use of two different methods.
The definition of POI is amenorrhea for ≥4 months with two serum follicle-stimulating hormone (FSH) levels in the menopausal range (obtained at least 1 month apart), or amenorrhea for ≥12 months before the age of 40 years. This definition is used in clinical practice, and therefore will be measured for pre-and perimenopausal women who visit the Survivorship Care Clinic.

However, some women were treated in the 1970s or 1980s, and developed POI already in the 1990s. In this case, measurement of FSH levels is not feasible as they have already been postmenopausal for many years. For these women, we therefore have to rely on 1. Their medical records stating the menopausal age, and 2. The self-reported menopausal age from the patient questionnaire. As we know menopausal age from the medical records for the majority of women, and from previous questionnaires mailed in the 1990s-2000s, recall bias is only a minor problem.
We added a sentence to the first paragraph of the Methods section Study parameters and data collection.

Added sentence: In case a woman has already been postmenopausal for many years at study enrolment, POI is defined as the date of or age at last menstruation. Because we performed earlier studies on POI, for the majority of women we already know their menopausal status and age, either from the medical records or from questionnaires sent in the 1990s-2000s. For the remaining women these data will be abstracted from the medical records and/or obtained through the patient questionnaire.

9. Pages 10, 11, Table 1: Asking patients to remember their last menstrual period and menopausal age, or medical documentation of these dates would be fraught with problems if the proposed study were performed in the US. Perhaps, in Netherlands it is different. If so, the protocol should state what is found in the medical records concerning menstrual function and age of menopause.

Reply: In principle we agree with the reviewer that it can be difficult to obtain data on menopausal age from interviews many years later. However, in our specific study setting we fortunately can make use of data available from previous studies in which we abstracted data on menopausal age from medical records and/or sent questionnaires on reproductive factors to female HL survivors. We refer to our answer to point 8.
Furthermore, in a previous study1 we compared medical record data on menopausal age with questionnaire data and observed agreement within a range of 2 years for the large majority of women.

1 Reference: Krul IM, Opstal-van Winden AWJ, Aleman BMP, et al. Breast Cancer Risk After Radiation Therapy for Hodgkin Lymphoma: Influence of Gonadal Hormone Exposure. Int J Radiat Oncol Biol Phys 2017;99(4):843-53 doi: 10.1016/j.ijrobp.2017.07.016.

10. If the design of the study was a prospective cohort design, perhaps you could make a case for the many biomarkers. But in observational cross-sectional design, many of the biomarkers are either not useful in answering the primary hypotheses or the data is conflicting on the individual biomarker. For example, a one-time assessment of bone turnover markers, CRP, NT-pro BNP is unnecessary. Also, the EKG and ECHO are not needed if you plan on performing the CCTA and IMT.

Reply: The SOPHIA-study is embedded in the Survivorship Care Clinics, which provide routine medical care for HL survivors. NT-pro-BNP, heart echo and EKG are part of routine medical care. Although these medical tests are not performed for research purposes, we will be able to include them in our analyses. With the patient's permission, we solely abstract the data from the medical records. Which medical tests are performed for routine care and which for research purposes depends on the treatment a patent received and is illustrated in detail in Figure 1 of the manuscript.

We are interested in possible differences in the main outcomes between HL survivors who developed POI and HL survivors who did not. These analyses can be adjusted for time since HL treatment and time since POI and are therefore still valuable despite the fact of a one-time measurement.

Of course, we agree that long-term follow-up leads to much more insight into this unique study population. Therefore, we intend to acquire additional funding in order to follow the study population longitudinally and to examine changes in risk factor and outcomes over time. We stated this in the original manuscript in the Methods section Design and study population. We have now placed more emphasis on the longitudinal follow-up of the study population by adding a sentence to the end of the Discussion section.

Added sentence: Finally, prospective follow-up of the study population will provide insight into longitudinal changes in risk factors and study outcomes.

11. Although the neurocognitive measures are justified, the problem is you don't know what the baseline pre-treatment neurocognitive function was. For example, studies that suggest neurocognitive in breast cancer patients pre-chemotherapy is abnormal in some domains.

Reply: Although we agree that the availability of baseline data would be an asset to our study, the practical situation in this clinical setting is that most of our patients were treated quite a while ago when neurocognitive measures were not considered. In the absence of baseline data we can still assess differences in mean scores of groups with and without POI, and address our study questions regarding neurocognitive outcomes. If there is an effect of POI, we certainly expect to see this at the group level.

12. Page 13, lines 1-15, the paragraph under medical records: Changing treatment practices, which authors acknowledge, could have a significant effect on outcome variables, but it is not likely the study is adequately powered to account the de-escalation of treatment that has occurred over time.

Reply: Our cohort is unique in that it has a high prevalence of POI in women treated before 1985. Indeed, most contemporary treatments are less gonadotoxic, but many patients with high stage HL still receive gonadotoxic treatment. Therefore, our aim is to investigate whether POI is associated with our outcomes of interest and not to investigate the effect of treatments over time. As often in late effects research, we will only be able to directly examine the effect of modern treatments in the future.

13. Page 14, lines 1-13, the paragraph under blood samples. Banking blood samples are great idea. Running these biomarkers in one-time assessment is not.

Reply: Cohort studies take time, therefore we feel that cross-sectional measurements are important as a starting point. As mentioned at question 10, we agree of course that long-term follow-up leads to much more insight into this unique study population. As this study is embedded in the Survivorship Care Clinics, eligible patients will be followed through clinical care where permission is asked to store future blood samples as well. Furthermore, we intend to acquire additional funding in order to follow the study population longitudinally and to examine changes in risk factor and outcomes over time. We stated this in the original manuscript in the Methods section Design and study population. To further clarify this point we added a sentence to this section.

Added sentence: Moreover, eligible women will be followed through clinical care, where permission is asked to store future blood samples as well.

14. Page 16, lines 2: The analysis of T-scores, not bone density, is the important variable. Everyone loses bone, but the T-scores are predictive fractures and is a more meaningful endpoint.

Reply: We will perform a DEXA-scan in combination with an instant vertebral assessment. During this examination, data on BMD, as well as T-scores and vertebral fractures become available. We agree that T-scores and vertebral fractures are the most important outcome measures, but we feel like it is important to be complete and show all variables related to bone health. The T-scores as outcome measure are mentioned in table 1.

15. Page 16, lines 31-60, the paragraph under statistical analyses: Should the p-values be more rigorous (lower) because of the potential multiple comparisons problem? Or is the sample size large enough to perform all of the proposed analyses?

Reply: We performed sample size calculations for each main outcome of interest separately in collaboration with a senior statistician (Dr. Michael Hauptmann). Therefore, we expect to be able to perform all the proposed analyses with sufficient power unless indicated in the manuscript. Indeed, the analyses of HRT use may lack power.
We do not think we need to account for multiple comparisons as our study is not a fishing expedition but based on clear hypotheses for each of the primary outcomes.

In summary, this study proposal strikes me as too ambitious with a multitude of questions on a relatively small, selected group of survivors. How is POF defined is a critical issue and the results are subject to recall bias or lack of documentation. If this not the case, then the authors should state the accuracy of the data sources determining POF. The protocol includes unnecessary testing, and the failure to separate the primary hypotheses from secondary exploratory ones. Also, this protocol would benefit from an independent statistical review to ensure they have an adequate patient population to answer the primary hypotheses.

Reply: We addressed all the above concerns extensively above and we hope we have answered all concerns in a satisfactory manner.

VERSION 2 – REVIEW

REVIEWER	Charles l. Shapiro Icahn School of Medicine at Mt Sinai, One Gustave Levy Place, Box 1079, NY, NY 10079
REVIEW RETURNED	26-Dec-2017

GENERAL COMMENTS	The authors state in the revised protocol that the secondary aims 2 and 3 will examine whether long-term effects differ between women that develop POI >1 or > 1 year if there is sufficient power (secondary aim 2), and to investigate the type and timing of HRT on all outcomes (secondary aim 3) if there is sufficient power, respectively. It is very unusual to state secondary aims with the caveat "if there is sufficient power." I would consider leaving these off. In the response to reviewers, page 19, they state they expect to perform all the primary endpoint analyses with sufficient power without adjusting the p values for multiple comparisons. They further state since they have a hypothesis for each of the primary endpoints, and their study is not a "fishing expedition" they think they do not have to account for multiple comparisons. But these endpoints are likely not independent each other. Furthermore, the statistical analysis includes a plan to analyze the age at treatment and age at developing POI and adjusting all analyses for cofounders including HL treatment, lifestyle factors, reproductive factors, climacteric symptoms, and medications when appropriate. I am not a statistician, but seems to me with 300 patients total, it may very difficult to all perform these analyses without getting results that either have wide confidence intervals, or inadequate power. This is why I request that this protocol undergo independent statistical review.

REVIEWER	James F. Troendle National Institutes of Health, National Heart, Lung, and Blood Institute, Division of Cardiovascular Sciences, Office of Biostatistics Research, United States
REVIEW RETURNED	25-Jan-2018

GENERAL COMMENTS	The paper describes a cohort study to look at differences in POI and non-POI survivors of Hodgkin lymphoma. The authors seem to want to find out if there are differences in BMD, cardiovascular status, neurocognitive function, or QoL by POI status. To do this they will measure these outcomes on a sample of Hodgkin lymphoma survivors. The problem comes in interpreting these results and determining if the results have clinical meaning. Because of possible survivorship bias, any finding of association (from a cross-sectional study) between POI status and BMD for example, is potentially misleading. In other words, one could find a lowering of BMD in POI women versus non-POI women, when in fact the opposite might be found in a longitudinal study. A simple theoretical example illustrates this point. Suppose, the group of women destined to be POI, is evenly divided into two sub-groups. One subgroup has high BMD values (compared to non-POI women), but extremely high incidence of death. The other subgroup has a moderately lower BMD (again compared to non-POI women) but similar mortality to non-POI women. The cross-sectional study will mostly find POI women in the second sub-group and thus conclude that POI is associated with lower BMD. However, the POI women overall have higher BMD in this example, as a longitudinal study would rightly find. The authors point out that they could also look at survival in the entire registry to see if there are differences in survival – and indeed this is possibly the best use of the registry. But finding or not finding a difference in survival will not remove the problem with interpretation of the cross-sectional findings on BMD etc. You will not know to what degree the cross-sectional finding was influenced by survivorship.

REVIEWER	Dr. Diana Tichy German Cancer Research Center, Division of Biostatistics, Heidelberg, Germany
REVIEW RETURNED	25-Feb-2018

GENERAL COMMENTS	1) Cohort of HL survivors includes patients treated between 1965 and 1985 where patients with natural POI are diagnosed between 1992 and 2012. Comparison of treatment induced POI and natural POI may be influenced by shifted age at menopause entry during the last 50 years. How is this aspect planned to be considered in the analysis? 2) You report, that there exists reverse causality with respect to CVD and POI risk. But isn`t there simply a correlation between CVD and POI caused by weight, cholesterol and blood pressure known as risk factors for both CVD and POI? 3) Sentence in line 48/49 is unclear. 4) With respect to line 30: which "outcomes of HL survivors" are meant? Is it just a descriptive analysis of the two cohorts of HL survivors developed POI and those did not? It should be mentioned that this is done with focus on explanatory analysis. 5) Long-term effects of treatment-induced POI are of primary interest of this cross-sectional study. These effects will be examined using multivariate regression. So far the application of cox regression

	models is unclear as described in line 39. Please explain in more detail which time to event is analysed. Further it is mentioned that “all analysis will be adjusted for confounders …”. One should consider the set of meaningful confounders for every analysis separately to avoid reduction in power. 6) The comparison of CVD status between HL survivors with POI and women with natural POI might be preceded by a propensity score analysis. Doing so one is able to reduce the bias in estimating the effect of treatment caused POI compared to natural POI. To investigate the difference in risk of CVD (time to CVD) the application of cox regression analysis might be the correct analysis approach, if the effect of age on CVD is carefully considered. 7) From my opinion: the main challenge with respect on the type of this cross sectional study is the following: an increased risk for CVD is expected under HL treatment caused POI compared to naturally induced POI. To investigate the difference in CVD risk, a cohort of HL survivors treated between 1965 and 1992 is compared with a cohort of women with natural POI diagnosed between 1992 and 2012. Hence women in the HL cohort are older on average and therefore have higher risk for CVD. The study can be conducted as described but one should adjust for/ be aware of age and time effects when investigating the effects of treatment caused POI on CVD compared to naturally induced POI. 8) The abstract should already contain an outline of the study`s strengths and limitations.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2

Reviewer Name
Charles l. Shapiro

Institution and Country
Icahn School of Medicine at Mt Sinai
One Gustave Levy Place
Box 1079
NY, NY 10079

Please state any competing interests or state ‘None declared’:
None declared

1. The authors state in the revised protocol that the secondary aims 2 and 3 will examine whether long-term effects differ between women that develop POI >1 or > 1 year if there is sufficient power (secondary aim 2), and to investigate the type and timing of HRT on all outcomes (secondary aim 3) if there is sufficient power, respectively. It is very unusual to state secondary aims with the caveat "if there is sufficient power." I would consider leaving these off.
Reply: thank you for your comment. This caveat was added to our revised manuscript as we think that these research questions are important to explore despite the fact that we may lack power. Since these analyses will be exploratory they may give direction to further studies. We agree with the reviewer that the exploratory character of the analyses addressing these secondary aims should be more emphasized. We therefore adjusted the sentence and switched the order of the aims (page number 7).

Adjusted sentence: “In addition, we will perform exploratory analyses to examine potential differences between subgroups regarding acute (<1 year after HL treatment) and more gradually (≥ 1 year after HL treatment) developed POI and to explore the effects of type and timing of HRT on all outcomes.”

2. In the response to reviewers, page 9, they state they expect to perform all the primary endpoint analyses with sufficient power without adjusting the p values for multiple comparisons. They further state since they have a hypothesis for each of the primary endpoints, and their study is not a "fishing expedition" they think they do not have to account for multiple comparisons. But these endpoints are likely not independent each other.
The reviewer raises an important point. The study as presented is powered for each outcome separately. However, conclusions about associations between POI and the different outcomes will not be based on a single test but on how plausible a true association is given the results of the various analyses specified. Instead of formally adjusting p-values for multiple comparisons, we will consider the possibility of a type 1 error in our interpretation of the results for each outcome. This point is added to the manuscript in the power calculation section, page number 16.

Added sentences: However, conclusions about associations between POI and the different outcomes will not be based on a single test but on how plausible a true association is given the results of the various analyses specified. Instead of formally adjusting p-values for multiple comparisons, we will consider the possibility of a type 1 error in our interpretation of the results for each outcome.

3. Furthermore, the statistical analysis includes a plan to analyse the age at treatment and age at developing POI and adjusting all analyses for cofounders including HL treatment, lifestyle factors, reproductive factors, climacteric symptoms, and medications when appropriate. I am not a statistician, but seems to me with 300 patients total, it may very difficult to all perform these analyses without getting results that either have wide confidence intervals, or inadequate power.
This is why I request that this protocol undergo independent statistical review.
We clarify our handling of confounding factors as follows: one by one, potential confounders (HL treatment, lifestyle factors, reproductive factors, climacteric symptoms, and medications) will be added to models with POI for the different outcomes to determine which of those are confounding the POI effect. Only those will be retained, and we expect these to be few. However, if the number would be large, we will use propensity score analyses instead.

Age at developing POI will be considered as a modifying factor, i.e., if POI is truly associated with an outcome, we will explore whether there is a stronger effect among women younger at POI versus those without POI. Therefore, since no conclusions are drawn on the independent effects of these additional factors, they do not necessarily reduce the power of our study. On the other side, we agree with the reviewer that, in case of multiple confounders, 300 patients will not allow strong conclusions on several outcomes unless effects are strong. We would like to emphasise that the SOPHIA-study is the first study addressing multiple outcomes (BMD, CVD and neurocognitive function) after POI in cancer survivors. Therefore, independent confirmation of our findings will certainly be needed.

We adjusted the sentence in the Statistical analyses section to make our handling of the confounders and the potential use of propensity scores more clear.

Old sentence: All analyses will be adjusted for confounders (HL treatment regimen, lifestyle factors, reproductive factors, climacteric symptoms, medication) where applicable.

New sentence: Potential confounders (HL treatment, lifestyle factors, reproductive factors, climacteric symptoms, and medications) will be added one by one to models with POI and the different main outcome variables to determine whether they are confounding the POI effect. Propensity score analyses will be used instead of adjustment if the number of confounders is large.

Reviewer: 3

Reviewer Name
James F. Troendle

Institution and Country
National Institutes of Health
National Heart, Lung, and Blood Institute Division of Cardiovascular Sciences Office of Biostatistics Research United States

Please state any competing interests or state 'None declared':
None declared

The paper describes a cohort study to look at differences in POI and non-POI survivors of Hodgkin lymphoma. The authors seem to want to find out if there are differences in BMD, cardiovascular status, neurocognitive function, or QoL by POI status. To do this they will measure these outcomes on a sample of Hodgkin lymphoma survivors. The problem comes in interpreting these results and determining if the results have clinical meaning. Because of possible survivorship bias, any finding of association (from a cross-sectional study) between POI status and BMD for example, is potentially misleading.
In other words, one could find a lowering of BMD in POI women versus non-POI women, when in fact the opposite might be found in a longitudinal study. A simple theoretical example illustrates this point. Suppose, the group of women destined to be POI, is evenly divided into two sub-groups. One subgroup has high BMD values (compared to non-POI women), but extremely high incidence of death. The other subgroup has a moderately lower BMD (again compared to non-POI women) but similar mortality to non-POI women. The cross-sectional study will mostly find POI women in the second sub-group and thus conclude that POI is associated with lower BMD. However, the POI women overall have higher BMD in this example, as a longitudinal study would rightly find. The authors point out that they could also look at survival in the entire registry to see if there are differences in survival – and indeed this is possibly the best use of the registry. But finding or not finding a difference in survival will not remove the problem with interpretation of the cross-sectional findings on BMD etc. You will not know to what degree the cross-sectional finding was influenced by survivorship.
Reply: We agree that a study like ours misses some patients who die before inclusion, and this can potentially influence the results. However, as far as we know, the outcomes considered in our study are not or only mildly (and certainly not strongly) associated with death rates among 8-year survivors of Hodgkin lymphoma under the age of 75 years. Therefore, survivorship bias will likely be minor. In contrast with most cross-sectional studies, our study is nested within a well characterized cohort, for which we have complete information on CVD outcomes, second malignancies, vital status and cause of death. An evaluation of differences in disease risks using our entire cohort will shed further light on survivorship bias. This will enable us to quantify any survivorship bias and to adequately interpret the strengths of our results. We added a sentence to the Discussion section on page 18.

Added sentence: This also allows us to evaluate potential differences in disease risks between participants and our entire cohort in order to quantify any survivorship bias and to adequately interpret the strengths of our results.

<u>**Reviewer: 4**</u>

Reviewer Name
Dr. Diana Tichy

Institution and Country
German Cancer Research Center, Division of Biostatistics, Heidelberg, Germany

Please state any competing interests or state 'None declared':
None declared

1) Cohort of HL survivors includes patients treated between 1965 and 1985 where patients with natural POI are diagnosed between 1992 and 2012. Comparison of treatment induced POI and natural POI may be influenced by shifted age at menopause entry during the last 50 years. How is this aspect planned to be considered in the analysis?

Thank you for your comments. We do not think this aspect is a problem in our analyses. There will be overlap in year at developing POI in the two groups as, for example, patients treated at age 20 for Hodgkin lymphoma in 1980, may experience POI more than 10 years after HL treatment (between 1990 and 2000). Second, the secular trend in menopausal age is weak and actually a bit controversial due to the use of hormone replacement therapy. Furthermore, the focus of our analyses is to examine late effects of POI and not menopause at later ages.

2. You report, that there exists reverse causality with respect to CVD and POI risk. But isn`t there simply a correlation between CVD and POI caused by weight, cholesterol and blood pressure known as risk factors for both CVD and POI?
Reply: in our manuscript on page 5 we mention three hypotheses that were postulated in the literature. The first hypothesis is that POI causes CVD through ovarian hormone deficiency. The second hypothesis is the reverse causality hypothesis, and the third one is that several risk factors underlie both CVD and POI. The reviewer refers to this last hypothesis, which might be true. However, to date there is no evidence yet and we would like to investigate these hypotheses in our study.

3. Sentence in line 48/49 is unclear.
Reply: In the absence of a page number, we believe the reviewer means the sentence on page 5, which has now been clarified to read:

Old sentence: A direct comparison of CVD risk between HL survivors and women with natural POI might provide new insights, as POI among HL survivors is induced by treatment (exogenous factors) instead of natural early depletion of the primordial follicle pool (endogenous factors).

New sentence: Since POI among HL survivors is induced by exogenous factors (i.e. HL treatment) rather than by endogenous factors (i.e. natural early depletion of the primordial follicle pool) occurring in women with natural POI, a direct comparison between HL survivors and women with natural POI might provide new insights into the association between POI and CVD.

4. With respect to line 30: which “outcomes of HL survivors” are meant? Is it just a descriptive analysis of the two cohorts of HL survivors developed POI and those did not? It should be mentioned that this is done with focus on explanatory analysis.
Reply: these are indeed descriptive analyses and in this sentence, we mean the characteristics of HL survivors. We adjusted the sentence.

Old sentence: Outcomes of female HL survivors who developed POI will be compared with those of female HL survivors who did not by using chi square tests or Fisher's exact tests (categorical variables) and two tailed t-tests (continuous variables) after appropriate transformation, if necessary.

New sentence: Characteristics of female HL survivors who developed POI will be compared with those of female HL survivors who did not by using chi square tests or Fisher's exact tests (categorical variables) and two tailed t-tests (continuous variables) after appropriate transformation, if necessary.

5. Long-term effects of treatment-induced POI are of primary interest of this cross-sectional study. These effects will be examined using multivariate regression. So far the application of cox regression models is unclear as described in line 39. Please explain in more detail which time to event is analysed.
Reply: The Cox regression models with age as a time scale will only be used to examine the independent effect of age at HL treatment on time to POI. This is a secondary analysis to create additional information on the determinants of POI in our study. We now see that this sentence was not completely correctly formulated. To clarify that time to POI is the outcome in this analysis, we corrected the sentence about the Cox regression in the statistical analysis section:

Old sentence: Cox regression models with age as a time scale will be used to examine the independent effect of age at HL treatment and age at developing POI.

New sentence: Cox regression models with age as a time scale will be used to examine the independent effect of age at HL treatment on age at developing POI.

6. Further it is mentioned that "all analysis will be adjusted for confounders …". One should consider the set of meaningful confounders for every analysis separately to avoid reduction in power.
We agree with the reviewer. For each analysis, confounders (HL treatment regimen, lifestyle factors, reproductive factors, climacteric symptoms, medication) will be identified and appropriately accounted for. Propensity score analyses will be used instead of adjustment if the number of confounders is large. To clarify this in our manuscript, we adjusted a sentence and added one sentence in the Statistical analyses section.

Adjusted sentence: Potential confounders (HL treatment, lifestyle factors, reproductive factors, climacteric symptoms, and medications) will be one by one added to models with POI and the different main outcome variables to determine whether they are confounding the POI effect.

Added sentence: Propensity score analyses will be used instead of adjustment if the number of confounders is large

7. The comparison of CVD status between HL survivors with POI and women with natural POI might be preceded by a propensity score analysis. Doing so one is able to reduce the bias in estimating the effect of treatment caused POI compared to natural POI.
We agree with the reviewer and have added a general comment about the use of propensity score analyses to control confounding.
Added sentence: Propensity score analyses will be used instead of adjustment if the number of confounders is large.

8. To investigate the difference in risk of CVD (time to CVD) the application of cox regression analysis might be the correct analysis approach, if the effect of age on CVD is carefully considered.
As explained above (comment 5), the Cox analysis will be on time to POI. This has been clarified in the manuscript in the Statistical analyses section. We do not observe time to CVD, as this is a cross-sectional study.

9. From my opinion: the main challenge with respect on the type of this cross sectional study is the following: an increased risk for CVD is expected under HL treatment caused POI compared to naturally induced POI. To investigate the difference in CVD risk, a cohort of HL survivors treated between 1965 and 1992 is compared with a cohort of women with natural POI diagnosed between 1992 and 2012. Hence women in the HL cohort are older on average and therefore have higher risk for CVD. The study can be conducted as described but one should adjust for/ be aware of age and time effects when investigating the effects of treatment caused POI on CVD compared to naturally induced POI.
Reply: we agree with the reviewer that age effects are important and we will adjust for age in the analyses. Furthermore, it should be noted that HL survivors are treated between 1965 and 2000, but that many develop POI several or even many years after treatment, so there will be overlap in year at POI between the two groups.

10. The abstract should already contain an outline of the study`s strengths and limitations.
We already mentioned the strengths and limitations in our original abstract on page 2.

VERSION 3 – REVIEW

REVIEWER	Dr. Diana Tichy German Cancer Research Center Heidelberg, Germany
REVIEW RETURNED	25-Apr-2018

GENERAL COMMENTS	The authors replied to my recent comments sufficiently, I have no further objections.