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)	Cohort-based association study of germline genetic variants with acute and chronic health complications of childhood cancer and its treatment: Genetic risks for childhood cancer complications Switzerland (GECCOS) study protocol
AUTHORS	Waespe, Nicolas; Strebel, Sven; Nava, Tiago; Uppugunduri, Chakradhara; Marino, Denis; Mattiello, Veneranda; Otth, Maria; Gumy Pause, Fabienne; Von Bueren, Andre; Baleydier, Frederic; Mader, Luzius; Spoerri, Adrian; Kuehni, Claudia; Ansari, Marc

VERSION 1 – REVIEW

REVIEWER	Lindsay Morton National Institutes of Health
REVIEW RETURNED	14-Jul-2021

GENERAL COMMENTS	Page 6, Lines 109-126: The first paragraph of the introduction could provide either a more comprehensive reference list or use of recent review papers to give the reader more context. For example, key papers on chronic health conditions (e.g., PMID 30416076) and mortality (e.g., PMID 33002115), or reviews could be used (e.g., from the Dec 2020 special issue of Pediatr Clin North Am). Page 7, Lines 151-154: Suggest acknowledgement of other large cohorts with DNA (e.g., the St. Jude Lifetime Cohort, the French Childhood Cancer Survivor Study) and also the use of case-control designs with DNA to address specific outcomes (e.g., PMID 30573632 in addition to other references already cited). General: Although some data are available on patients going back to 1976, the availability of biospecimens for each patient and that proportion that have consented are unclear. The text and figures would benefit from description of the biospecimen availability and consent proportions as well as discussions of potential issues related to survivor bias.

REVIEWER	Nan Song Chungbuk National University, College of Pharmacy
REVIEW RETURNED	22-Oct-2021

GENERAL COMMENTS	This article seems to have described GECCOS study protocol well as aspect of study concepts and study designs. But several issues should be described more before publication to improve the quality of protocol paper. Major concerns 1. In the last paragraph of the 1. Introduction section, author mentioned previous studies' limitations, such as small sample size,

	heterogenous cohorts with various treatment exposures, inconsistent outcome assessments, investigation of insufficient health conditions. Did the current study overcome those limitations? Would you please add some description on previous studies' limitations and strengths of the current study in the discussion section? 2. It would be better to make supplementary table on description of data sources including SCCR, SCCSS, BISKIDS, and BaHOP in order to make readers understand easily. 3. In 3.3 Study population section, author considered people who were diagnosed with Langerhans cell histiocytosis (LCH) as eligible for this study. There needs rationale of selection of those people. 4. Is there any follow-up information? 5. The contents of study design (cohort, case-contro, etc.) in 3.3. Study population section would be better to move into 3.1. Study design. 6. Are there any lifestyle or envirionmental factors collected? 7. In 3.7. Data linkage, I am wondering whether author do not need unification or harmonization processes. 8. In 3.10. Power calculation, power could be calculated with sample size, possible effect sizes before study. Could you calculate possible powers of the study with some examples? 9. Validation and replication process should be specified. Minor concerns 1. The contents in 'Strengths and limitations of this study' section in ARTICLE SUMMARY on page 5 seems to indicate not strengths and limitations but summary of the study. 2. Inclusion and exclusion criteria should be in 3.6. Selection of participants section with N (%). The relevant contents seem to be described in the middle of 3.2. Data sources section. 3. Would you re-arrange sub-headings of the methods section? It would be better to connect 3.3 Study population and 3.6. Selection of participants. 4. There is a typo in the first paragraph of the 3.11. In silico and invtro analyses. ")"

VERSION 1 – AUTHOR RESPONSE

Dr. Lindsay Morton, National Institutes of Health
Comments to the Author:
Page 6, Lines 109-126: The first paragraph of the introduction could provide either a more comprehensive reference list or use of recent review papers to give the reader more context. For example, key papers on chronic health conditions (e.g., PMID 30416076) and mortality (e.g., PMID 33002115), or reviews could be used (e.g., from the Dec 2020 special issue of Pediatr Clin North Am).
Thank you for suggesting additional references. We added them to the introduction and adapted the text accordingly:
L. 131-134: "In more recent decades, chronic health conditions were reduced following treatment adaptations to reduce adverse events.[5] Mortality is significantly increased in survivors compared to the general population [6] and varies depending on the treatment exposure over time.[7]" […]
L. 165-167: "Replication of findings is necessary before allowing translation of these findings into clinical practise.[37] For many SPNs (such as thyroid cancer), no data is available.[38]"

Page 7, Lines 151-154: Suggest acknowledgement of other large cohorts with DNA (e.g., the St. Jude Lifetime Cohort, the French Childhood Cancer Survivor Study) and also the use of case-control

designs with DNA to address specific outcomes (e.g., PMID 30573632 in addition to other references already cited).
We added existing large cohorts and the use of case-control design in research projects. The modified section reads as follows:
L. 167-173: “To address the contribution of genetic risk variants in the development of late toxicities, large cancer survivor studies such as the Childhood Cancer Survivor Study (CCSS)[36] and SJLIFE cohort[39] in the US and the French childhood cancer survivor study for leukaemia (LEA Cohort)[40] are collecting DNA systematically to conduct genotype-phenotype analyses.” […]
L. 182-183: “For specific outcomes, case-control designs using samples from different cohorts have also been successfully used.[41]”

General: Although some data are available on patients going back to 1976, the availability of biospecimens for each patient and that proportion that have consented are unclear. The text and figures would benefit from description of the biospecimen availability and consent proportions as well as discussions of potential issues related to survivor bias.
So far, we completed a first collection of germline DNA in a subsample of the cohort. Roughly 50% of survivors participated by donating saliva samples. We added information on this first cohort that was recruited and the respective reference:
L. 381-383: “In a first recruitment effort, 928 childhood cancer survivors from the SCCR were asked to participate in germline DNA collection by home saliva collection and 463 (50%) participated.[52]”
We added a section discussing survivor bias and how we intend to counteract it:
L. 518-522: “Collection of germline DNA in survivors was done in a first subset of participants. The collection is subject to survivor bias and omission of patients who died before they could be invited to germline DNA collection. This would lead to selection of patients with less severe phenotypes. We included a second stream of collection of samples from newly diagnosed patients through participating hospitals to include all childhood cancer patients early after diagnosis.”

Reviewer: 2
Dr. Nan Song, St Jude Children's Research Hospital
Comments to the Author:
This article seems to have described GECCOS study protocol well as aspect of study concepts and study designs. But several issues should be described more before publication to improve the quality of protocol paper.

Major concerns
1. In the last paragraph of the 1. Introduction section, author mentioned previous studies’ limitations, such as small sample size, heterogenous cohorts with various treatment exposures, inconsistent outcome assessments, investigation of insufficient health conditions. Did the current study overcome those limitations? Would you please add some description on previous studies’ limitations and strengths of the current study in the discussion section?
We added the strengths of the current study in respect to the limitations of previous studies and expanded the discussion to address this topic:
L. 530-538: “To overcome the issue of many previous studies using small sample sizes, the GECCOS study recruits participants from the nationwide SCCR with more than 8000 childhood cancer patients and survivors. This large base cohort will allow selection of specific treatment exposures to create homogenous samples for specific genotype-phenotype associations. Several studies are planned or ongoing to assess long-term complications in a standardized way in Switzerland (e.g. cardiac outcomes)[69] which will improve quality of outcome assessments that can be used for the GECCOS study. Outcomes for health conditions that have been less investigated, like pulmonary complications, are also being collected and will be used in the GECCOS study.[70]”

2. It would be better to make supplementary table on description of data sources including SCCR, SCCSS, BISKIDS, and BaHOP in order to make readers understand easily.
 Thank you for this suggestion. We added a summary table to the manuscript to clarify the datasets used: Supplementary Table 1.

3. In 3.3 Study population section, author considered people who were diagnosed with Langerhans cell histiocytosis (LCH) as eligible for this study. There needs rationale of selection of those people.
We added a justification for adding patients with Langerhans cell histiocytosis:
L. 278-281: “Langerhans cell histiocytosis has been included in the SCCR. As LCH shows clonal proliferation of immature cells with somatic activating gene mutations and the need for antineoplastic treatment in an important subset of patients, we included this entity in the GECCOS study.[48]”

4. Is there any follow-up information?
Please find this information in table 1 and supplementary table 3, which we updated to highlight the availability of this data. We added further clarification on follow-up information:
L. 257-264: “The SCCSS collects questionnaire-based information from survivors on self-reported health outcomes, sociodemographic information, and environmental exposures (such as smoking). Clinical data on chronic health conditions after childhood cancer from survivorship clinics and hospital records (e.g. audiograms, and lung function tests) will be extracted for GECCOS. The datasets used for the association analyses will contain collected follow-up information from medical records, self-reported outcomes, and functional outcome data from specific projects on long-term complications.”

5. The contents of study design (cohort, case-contro, etc.) in 3.3. Study population section would be better to move into 3.1. Study design.
We moved the study design information to the section 3.1 on study design:
L. 221-224: “Within sub-projects assessing specific outcomes (such as hearing loss), we will sample patients and survivors according to risk exposure (for a cohort design), or according to the outcome of interest (for a case-control or case-cohort design).”

6. Are there any lifestyle or envirionmental factors collected?
Thank you for pointing out these important covariates. We added information on environmental and lifestyle factors that are available for the study cohort in the manuscript. Please also see our reply to your comment 4:
L. 257-264: “The SCCSS collects questionnaire-based information from survivors on self-reported health outcomes, sociodemographic information, and environmental exposures (such as smoking). Clinical data on chronic health conditions after childhood cancer from survivorship clinics and hospital records (e.g. audiograms, and lung function tests) will be extracted for GECCOS. The datasets used for the association analyses will contain collected follow-up information from medical records, self-reported outcomes, and functional outcome data from specific projects on long-term complications.”
Supplementary Table 3 was updated to highlight the information on “Self-reported symptoms and environmental exposure”.

7. In 3.7. Data linkage, I am wondering whether author do not need unification or harmonization processes.
We will use datasets to perform the analyses with datasets that do not contain overlapping information and that are clearly allocated to specific patients. Clinical data on primary diagnosis, treatment exposure and some follow-up data is collected from the SCCR, self-reported outcome information and extracted data from medical records from SCCSS, and genetic samples and data from BaHOP/ BISKIDS:
L. 366-369: “Coded unique identifiers allow linkage of genotype data from the BISKIDS collection to phenotype data from the SCCR and SCCSS. The identifiers are securely stored in a separate trust

center database managed by a third party (SwissRDL – Swiss Medical Registries and Data Linkage, ISPM, University of Bern)."

8. In 3.10. Power calculation, power could be calculated with sample size, possible effect sizes before study. Could you calculate possible powers of the study with some examples?
We added an example for the study size calculation:
L. 435-439: "For the sub-project on hearing loss, we have collected germline DNA from 426 survivors. Data collection and cleaning for the outcome measure (audiograms) is ongoing. We calculated sufficient power to detect a variant with a relative risk of 2.5 in an exome-wide association analysis using a dominant model."

9. Validation and replication process should be specified.
We expanded this section to better illustrate our validation and replication process:
L. 462-474: "We will seek to validate variants identified by next-generation sequencing that were associated with the respective outcome of interest using a different method (e.g., Sanger sequencing or real-time PCR). After successful validation, we will seek to proceed to replication of identified variants. We will reach out to independent cohorts of childhood cancer patients and survivors containing similar outcome information as analysed in the primary dataset such as the SJLIFE cohort or the French LEA cohort.[39,40] We will assess the power to identify an association in the replication cohort (using minor allele frequency of the identified variant in the respective population, the suspected effect size and sample size of the cohort). If patient numbers are deemed insufficient, we will consider pooling of data from multiple cohorts.

Minor concerns
1. The contents in 'Strengths and limitations of this study' section in ARTICLE SUMMARY on page 5 seems to indicate not strengths and limitations but summary of the study.
We revised this section and specifically addressed strengths and limitations:
L. 90-105:
- "The strength of the Genetic Risks for Childhood Cancer Complications Switzerland (GECCOS) study is the recruitment of childhood cancer patients and survivors from the national population-based Swiss Childhood Cancer Registry (SCCR) with data from 8,163 participants.
- The SCCR contains an extensive dataset including demographic, treatment, outcome, follow-up, and survival information which is then used for genotype-phenotype association analyses.
- The germline DNA collection within the Germline DNA Biobank for Childhood Cancer and Blood Disorders (BISKIDS) allows storage of samples and sequencing data creating an increasing collection of genetic material and data for future use.
- While the cohort for patient recruitment is large, the population with a specific outcome of interest might be small for specific populations. This limitation will be counteracted by actively seeking international collaborations with pooling of available data. "

2. Inclusion and exclusion criteria should be in 3.6. Selection of participants section with N (%). The relevant contents seem to be described in the middle of 3.2. Data sources section.
We changed the data as suggested to the section 3.6, which became 3.4 due to your minor comment 3:
L. 286-312: "We will identify participants eligible for specific sub-studies with defined in- and exclusion criteria. As of December 2019, 13,029 patients were registered in the SCCR, of which 9,306 (69%) were still alive and 8,163 (63%) were Swiss residents and potentially eligible for participation in GECCOS. We will use information in the SCCR and SCCSS and assess availability of corresponding germline DNA samples or sequencing data from previously sequenced participants in BISKIDS (Figure 2). If clinical data is available for a sufficient number of participants but further genetic samples are needed, we will invite potential participants to contribute germline DNA samples to

BISKIDS for research. For collection of biological material within BISKIDS, we will use two pathways: (i) invitations to participate are sent out by the Childhood Cancer Research group at the University of Bern, consisting of germline DNA collection kits (predominantly using saliva samples or buccal swabs) with information on the biobanking project and associated research, and informed consents to the participant's home; (ii) participants are invited by healthcare staff in hospitals caring for childhood cancer patients and survivors. These potential participants and their legal representatives are informed of the project and written consent and germline DNA are collected during a medical visit already planned for their treatment or follow-up. All participants consent to have their germline DNA stored in the BISKIDS section of the BaHOP biobank and their health-related data to be used for genotype-phenotype association studies. All specific GECCOS sub-studies will be reviewed and approved by a national scientific committee and submitted to the responsible ethics committee as amendment to the main GECCOS protocol, insofar as the applicable law requires authorization."

3. Would you re-arrange sub-headings of the methods section? It would be better to connect 3.3 Study population and 3.6. Selection of participants.
We rearranged the sub-headings as suggested and 3.6 became 3.4 following sub-heading 3.3.

4. There is a typo in the first paragraph of the 3.11. In silico and invtro analyses. ")"
Thank you for pointing this out. We adjusted the text:
"Examples of such models are clustering methods including similarity network fusion[42] and PEGASUS[64]" […]

VERSION 2 – REVIEW

REVIEWER	Lindsay Morton National Institutes of Health
REVIEW RETURNED	21-Dec-2021

GENERAL COMMENTS	The authors effectively addressed the reviewer comments.