PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

Title (Provisional)

Protocol for the Paediatric Personalized Research Network Switzerland (SwissPedHealth): A Joint Paediatric National Data Stream

Authors

Mozun, Rebeca; Belle, Fabiën N; Agostini, Andrea; Baumgartner, Matthias R; Fellay, Jacques; Forrest, Christopher; Froese, D Sean; Giannoni, Eric; Goetze, Sandra; Hofmann, Kathrin; Latzin, Philipp; Lauener, Roger; Martin Necker, Aurélie; Ormond, Kelly; Pachlopnik Schmid, Jana; Pedrioli, Patrick G A; Posfay-Barbe, Klara Maria; Rauch, Anita; M. Schulzke, Sven; Stocker, Martin; Spycher, Ben D; Vayena, Effy; Welzel, Tatjana; Zamboni, Nicola; Vogt, Julia E; Schlapbach, Luregn J; Bielicki , Julia Anna; Kuehni, Claudia

---

VERSION 1 - REVIEW

Reviewer	**1**
Name	**Goldstein, Neal D.**
Affiliation	**Drexel University**
Date	**14-Aug-2024**
COI	**None**

This protocol is a description of the SwissPedHealth intended to centralize collection of routine clinical data on the entire pediatric population of Switzerland, thus enabling research use of these data for public and clinical health questions. Overall the protocol is well written and logically organized. There is a lot of detail on how the data will be collected and the governance of SwissPedHealth but less on how this can be made a useful resource for researchers.

Some questions I would like to see addressed in a revision include:

1) How is data quality being handled (i.e. accuracy of EHR data, missing data, representativeness)? What level of error/missingness is deemed acceptable?

2) How is the use of and productivity from SwissPedsHealth being assessed (projects, citations, grants using these data, etc.)?

3) There appears to be a lack of analytic plans for the nested projects and a similar lack of detail for the lighthouse project. Please provide this detail, as well as define what is meant by “machine learning” as there are many approaches one could take.

4) How will researchers access the data in SwissPedHealth? Is it going to be a virtual de-identified environment with common statistical software/analytic tools pre-loaded? Can data be downloaded? What is the process for researchers to request access to SwissPedHealth? How is privacy and security of the data being ensured?

5) A lot of the information I would expect to see in a protocol – as indicated in the RECORD template – is labeled ‘NA’ presumably because data have not been analyzed. Defining analytic/statistical plans, variables, data sources, bias, and so on should be done a priori as part of the methods and should ad hoc once data are collected.

Reviewer	**2**
Name	**Hashmi, S. Shahrukh**
Affiliation	**University of Texas Health Science Center Houston**
Date	**29-Aug-2024**
COI	**None**

This manuscript details the protocol for SwissPedHealth, a data stream and infrastructure to collect routine pediatric clinical data in a standardized manner and make it available for research and health-policy purposes. The manuscript is well-written and detailed and adequately describes the protocol in place, how the data would be used, plans for the future, processes for review and feedback, and strengths and limitations.

The only question that wasn't answered for this reviewer is as follows:

The authors reported that for the multi-omics data (as part of the lighthouse project), the cohort of children will be identified and along with the patients' data and bio-samples, consents will be obtained for mucosal swabs from parents and siblings. Are these open-ended consents covering various future studies or are they more limited and study-specific? Do patients/families have the option to opt-out or opt-in of specific study types?

VERSION 1 - AUTHOR RESPONSE

Reviewer: 1
Dr. Neal D. Goldstein, Drexel University

Comments to the Author:
This protocol is a description of the SwissPedHealth intended to centralize collection of routine clinical data on the entire pediatric population of Switzerland, thus enabling research use of these data for

public and clinical health questions. Overall the protocol is well written and logically organized. There is a lot of detail on how the data will be collected and the governance of SwissPedHealth but less on how this can be made a useful resource for researchers.

Some questions I would like to see addressed in a revision include:

1) How is data quality being handled (i.e. accuracy of EHR data, missing data, representativeness)? What level of error/missingness is deemed acceptable?

Reply: We have now included details on the central data quality assessment.

Page 10, methods and analysis section, central data management:

> *"Data quality assessment*
>
> *The central data quality assessment includes the evaluation of data validity, accuracy, completeness, consistency, timeliness, and integrity on the NDS B-space, both individually for each CDW and across CDWs to investigate potential discrepancies between data providers. The output of this report will be fed-back to the CDWs for clarification and checks as necessary to iteratively improve data quality on the NDS B-space. Data will be transferred together with a data quality report to the project specific workspaces, where researchers will perform additional data quality assessments tailored to project needs. In particular, one of the nested projects is specifically designed to assess data quality completeness, representativeness and accuracy by comparing the datasets for children diagnosed with cancer against an external reference standard: the national Childhood Cancer Registry."*

Data quality handling will be discussed further in detail in future more technical manuscripts specific for the nested and lighthouse projects. The acceptable level of missing data will vary based on the specific requirements of the project and the nature of the variables involved. Each case will be evaluated individually to ensure alignment with project goals and data quality standards.

2) How is the use of and productivity from SwissPedsHealth being assessed (projects, citations, grants using these data, etc.)?

Reply: Thank you for this remark. We have added a section describing the use and productivity from SwissPedHealth.

Page 16, methods and analysis section:

> *"Use and productivity assessment plan for SwissPedHealth*
>
> *To assess the use of our data infrastructure, key metrics include the number of projects requesting data access and establishing regulatory documents, the volume of data requests for approved projects in BioMedIT workspaces, users per workspace, and data access frequency. Support requests from project-specific workspaces are analysed to identify and evaluate potential bottlenecks. Productivity of SwissPedHealth is assessed by monitoring research publications, citations, conference presentations, new collaborations, project plans, and grants. PPIE activity is recorded through minutes of focus groups, satisfaction surveys, number of projects and researchers using the planned PPIE toolbox and tracking outreach activities."*

3) There appears to be a lack of analytic plans for the nested projects and a similar lack of detail for the lighthouse project. Please provide this detail, as well as define what is meant by "machine learning" as there are many approaches one could take.

Reply: We have added details on the machine learning approaches and methods for the lighthouse project. For the nested projects, specific full study protocols including detailed data analysis plans will be published separately.

Page 12, methods and analysis section, nested projects

> *"Study protocols for the nested projects, including detailed data analysis plans will be published separately. Furthermore, the nested projects will partner with international research networks, to identify best practices, compare performance across countries, and enhance the learning of paediatric health systems. Full detailed study protocols for the nested projects, including specific data analysis plans will be published separately."*

Page 14, methods and analysis section, lighthouse project

> *"Analysis outline for the lighthouse project*
>
> *Overall, to the aim is to comprehensively assess the impact of genetic determinants on disease and their relationships with RNA, protein, and other metabolites. To search for novel disease-causing variants both in phase 2 and 3, we will use an optimized workflow for automated DNA variant calling and annotation, providing ready-to-interpret reports. Rare variants will be prioritized based on: 1) the potentially damaging effect of the variant; 2) the possible known function of the corresponding gene in the clinical presentation and 3) the degree of purifying selection to which they are subjected. RNA-seq data analysis will be used for quantitative expression profiling and alternative splicing analysis. DNA variants, RNA transcripts, proteoforms, and metabolites will be mapped onto biological networks and pathways using tools.*
>
> *Joint analysis will include statistical genomics to detect enrichment of disease-causing variants within specific phenotypes. To investigate genotype-phenotype relationships, we will examine genetic determinants of disease and their impact on RNA, protein and metabolite level. To analyse rare (or combined rare and common) DNA variation, we will use the same methods as common variant analysis with the addition of several necessary protocols. These methods will be used to analyse both rare and common DNA variants simultaneously, and rare variants exclusively, while adjusting for relevant covariates and incorporating multi-variate and multi-category outcome models. Further, we will perform gene- and protein-based pathway collapse analysis, incorporating covariates, to summarize variant-level information within functional units or biological pathways.*
>
> *We will perform RNA-seq data analysis using regression models to assess the relationship between gene expression and phenotypic outcomes while adjusting for potential confounders. For proteomic and metabolite analysis, we will map proteoforms and metabolites onto biological networks and pathways using enrichment analysis methods and apply multivariate statistical methods to identify patterns and relationships between proteomic or metabolomic profiles and phenotypes. Throughout, the major outcome variables will consist of extreme phenotypes such as inborn errors of metabolism and inborn errors of immunity. Strict thresholds for multiple test correction will be applied based on the number of tests. Candidate causal determinants will be interpreted based on best practices for clinical genetic reporting and American College of Medical Genetics and Genomics (ACMG) recommendations.[41,42]*
>
> *We additionally leverage machine learning (ML) for the analysis of the multi-omics data. In the context of this study, patients are characterized by means of multiple representations, or views - clinical, genomic, transcriptomic, proteomic, and metabolomics - each critically relevant to the development of a rare disease. While previous works mostly used self-supervised approaches such as autoencoders,[43] we will explore multimodal and multitask learning to find an optimal subset of explanatory genes, and aim to develop novel machine learning based approaches to the analysis of multi-omic datasets.[44-49]"*

4) How will researchers access the data in SwissPedHealth? Is it going to be a virtual de-identified environment with common statistical software/analytic tools pre-loaded? Can data be downloaded? What is the process for researchers to request access to SwissPedHealth? How is privacy and security of the data being ensured?

Reply: Thank you for allowing us to expand and describe details on the IT environment for data access and processing, we have now added a new section in the methods.

Page 8, methods and analysis section, SwissPedHealth governance structure and regulatory framework:

> *"SwissPedHealth data access regulation*
>
> *Access to the NDS data is regulated through the SwissPedHealth infrastructure consortium agreement. When planning to carry out a project using data from the SwissPedHealth NDS, researchers will need to set up a data project consortium agreement using our template and involving all participating partners. This should include "internal" SwissPedHealth partners and "external" partners, i.e. those not part of the SwissPedHealth NDS Infrastructure Consortium. For the data project consortium agreement, projects should consider the SwissPedHealth general terms and conditions and, wherever possible, try to adopt them without or with minor deviations. This will facilitate the process of legal review and signing for all internal partners, since the general terms and conditions will have already been reviewed as part of the development and adoption process. If deviations are required, these will need to be specified in the data project consortium agreement, and should be highlighted during its legal review. The following additional documentation will be required to complete the data project consortium agreement: project funding application/proof of funding, study protocol and ethical approval. In case the latter is not available it should be added as soon as available, and the data project consortium agreement cannot be executed without evidence of ethical review and approval. All data project consortium agreements will need to incorporate a data transfer and use agreement and data transfer and processing agreement. These are included in the SwissPedHealth data project consortium agreement template. Projects need to set-up a specific space in the BioMedIT secure IT network for data access and processing. Research groups access their B-spaces through the BioMedIT Portal, where they manage users, encryption keys, and data transfers, while also accessing support and tools. Researchers use two-factor authentication for secure access, and only extract nonsensitive or aggregated results, keeping sensitive data within the platform. Adaptations to the agreements may be necessary in cases where the Data Processors are not one of the BioMedIT nodes (SENSA – Lausanne, SciCORE – Basel, SIS – Zurich). For the DTUA section, each project will need to describe and depict the planned data flow as well as data and metadata to be transferred. Details on the process to access the data, as well as on data availability and contact information will be made available through the SPHN metadata catalogue, as well as information on data sources and qualitative and quantitative metadata on concept availability in the NDS."*

5) A lot of the information I would expect to see in a protocol – as indicated in the RECORD template – is labeled 'NA' presumably because data have not been analyzed. Defining analytic/statistical plans, variables, data sources, bias, and so on should be done a priori as part of the methods and should ad hoc once data are collected.

Reply: This protocol manuscript aims to provide an overview of the entire SwissPedHealth national data stream, including presentation of the key design, infrastructure, governance, PPI elements as well as specific nested and lighthouse project. We believe therefore it is important to provide sufficient space to cover the project infrastructure, for better text readability and length considerations. In response to the above comment, we have now included more key details on the analysis plans within

the lighthouse project section. In addition to the lighthouse project, four nested projects are planned within SwissPedHealth, which cover a broad range of topics. Project specific manuscripts and protocols will discuss in detail statistical analysis including variables, data sources, handling of bias etc.

We now indicate in the methods that information on data sources, and metadata on data availability will be accessible through the SPHN metadata catalogue, and mention in the discussion representativeness as potential selection bias.

Page 9, Methods and analysis, SwissPedHealth data access regulation:

> *"Details on the process to access the data and contact information will be made available through the SPHN metadata catalogue, as well as information on data sources and qualitative and quantitative metadata on concept availability in the NDS. the NDS.*

Page 19, discussion section, strengths and limitations:

> *"Acceptance rates of General Consent or Informed Consents and regulatory allowances will affect representativeness of the study populations for projects conducted with SwissPedHealth."*

Reviewer: 2
Dr. S. Shahrukh Hashmi, University of Texas Health Science Center Houston

Comments to the Author:
This manuscript details the protocol for SwissPedHealth, a data stream and infrastructure to collect routine pediatric clinical data in a standardized manner and make it available for research and health-policy purposes. The manuscript is well-written and detailed and adequately describes the protocol in place, how the data would be used, plans for the future, processes for review and feedback, and strengths and limitations.

The only question that wasn't answered for this reviewer is as follows:
The authors reported that for the multi-omics data (as part of the lighthouse project), the cohort of children will be identified and along with the patients' data and bio-samples, consents will be obtained for mucosal swabs from parents and siblings. Are these open-ended consents covering various future studies or are they more limited and study-specific? Do patients/families have the option to opt-out or opt-in of specific study types?

Reply: Thank you for allowing us to clarify this. The informed consent form of the lighthouse project includes a declaration of consent for further use of (genetic) data and biological material in encrypted form for patients' blood samples as well as for buccal swab samples from their parents and healthy siblings. We now specify this in the text. While the consent form for further use is not specific for different study types, further use of data will require approval by the respective ethical committees.

Page 13, methods and analysis section, the lighthouse project:

> *"We also obtain consent for the encrypted use of patients' and relatives' (genetic) data and biological material for future research."*