

# 2024 - 2026 plans

Dylan Lawless

2024-05-28

## Contents

<b>1 Checklist</b>	<b>1</b>
<b>2 Review: 2023.06.01 -2024.06.01</b>	<b>3</b>
2.1 [Analysis package] ACMGuru . . . . .	3
2.2 [Analysis design] GenomeSwift . . . . .	3
2.3 [Analysis design] deepInfer . . . . .	3
2.4 [Manuscript - review] Statistical developments in variant set association testing . . . . .	4
2.5 [Manuscript - finished] Rare variants in infection response protein pathway associated with sepsis in children. . . . .	4
2.6 [Manuscript - finished] Genome-wide association study of pediatric sepsis. . . . .	5
2.7 [Manuscript - published] You AIn't using it right - artificial intelligence progress in allergy . .	5
2.8 [Grant application] Heidi Ras . . . . .	6
2.9 [Grant application] NCCR assist . . . . .	6
2.10 [Pipeline develop] Phase 1-2 variant analysis - ML and statistics datasets prep. . . . .	6
2.11 [Pipeline develop] Phase 1-2 statistical genomics and clinical genetics report prep. . . . .	7
2.12 [Concept develop] SwissPedHealth contribution to data models . . . . .	7
<b>3 Plans for 2025-2026</b>	<b>7</b>
3.1 Future Plan: Joint analysis of SwissPdHealth phases 1-3 . . . . .	7
3.2 Future Plan: Directly integrate genetic data concepts into clinical databases . . . . .	7
3.3 Future Plan: Publish our SwissPedHealth genomics system. . . . .	8
3.4 Future Plan: Publish joint genomics analysis of SwissPedHealth phase 1-3 . . . . .	8
3.5 Future Plan: Publish/collaborate on the join DNA, RNA, proteomics analysis of SwissPed-Health phase 1-3 (Sean, Daphné, Vito) . . . . .	8
3.6 Future Plan: Publish deepInfer - a continuously updating database of causal inference. . . .	8
3.7 Future needs: SciCORE environment . . . . .	9
3.8 Future needs: Research projects . . . . .	9
3.9 Future needs: Precision medicine . . . . .	9





---

## 1 Checklist




1. State of current projects and completion of SPSS-papers.
2. Feedback from you on:
  - The current setting you are working in, what works, what can we improve?
  - Your goals, vision for the next 2 years and discussing a plan on what is needed to make that “materialise”?
3. Any other feedback or priority you want to discuss?

## Pipeline modules - WGS

## Phase 1 – Testing ACMGuru package

- Prepare analysis-ready datasets 
- Qualifying variant selection 
- Tailored filtering 
- Result interpretation  (manual)

## Phase 2 – Testing pipeline automation

- GATK pipeline optimise 
- Storage and costs 
- ACMGuru reports  (template done)

### Phase 3 – Automated run

## Goals

## WGS turnaround

- Single sample turn-around time
- Joint cohort analysis time / cost

## Varian association tests

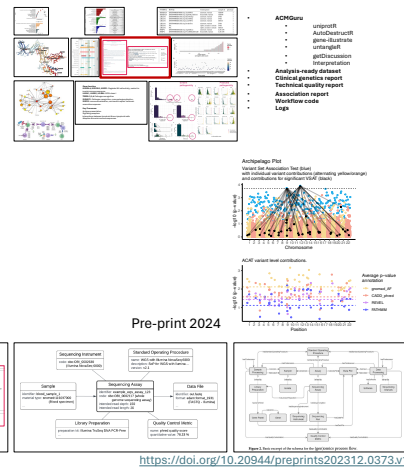
- Genetic association with outcome including multi-omic input

## SPHN RDF Schema

- Output directly

## Formal test

- genomebottle pangenome



## Future



## 2 Review: 2023.06.01 -2024.06.01

### 2.1 [Analysis package] ACMGuru

**Objective:** [ACMGuru is an R package](#) designed to facilitate the interpretation of genetic determinants of disease in genomic data. It applies extensive annotation and filtering based on the ACMG (American College of Medical Genetics and Genomics) and AMP (Association for Molecular Pathology) guideline standards.

**Basis:** The foundational document in this field is the joint consensus recommendation on the standards and guidelines for the interpretation of sequence variants by Richards et al. (2015, doi: 10.1038/gim.2015.30). ACMGuru is designed to implement the initial steps of our filtering protocol for the addition of ACMG-standardised labels to candidate causal variants, according to these standards. A number of additional advances in variant interpretation are implemented (2021, Pedersen, doi: 10.1038/s41525-021-00227-3)

**Research Plan:** ACMGuru enhances the interpretation of genomic data through automated annotation and filtering based on ACMG/AMP guidelines. The project will be developed in the following stages: **1. Implementation:** ACMGuru uses VEP plugins and their associated databases, as well as the dbNSFP plugin and database for SNVs, to apply annotation features and ACMG rules. **2. Annotation and Filtering:** Key features used for annotation include sample ID, gene symbols, genomic positions, allele frequencies, and various predictive scores (e.g., CADD, PolyPhen). **3. Visualization:** The package provide tools to visualise the results of filtering and annotation processes. **4. Validation:** Rigorous testing and validation processes ensures the accuracy and reliability of the annotations and interpretations provided by ACMGuru.

### 2.2 [Analysis design] GenomeSwift

**Objective:** The primary goal of GenomeSwift is to develop and deploy a comprehensive, automated software pipeline that enhances the speed and accuracy of genomic data analysis, particularly for diagnosing and treating genetic diseases in Kispi.

**Basis:** We have demonstrated and validated each of the individual components. We now merge them into a unified pipeline.

**Research Plan:** GenomeSwift will integrate several purpose-build tools (ProteoMCLustR, SkatRbrain, Archipelago, ACMGuru, and deepInferR), along with several industry-standards, into one unified platform to optimise genomic data processing from initial input to clinical interpretation. The project will employ robust statistical methodologies and simulation/validation processes to ensure reliability and clinical relevance.

### 2.3 [Analysis design] deepInferR

**Objective:** The primary goal of deepInferR is to develop a Bayesian framework for pre-calculating the probability of genetic variants causing specific diseases. This involves integrating known priors and estimated unknowns to provide robust disease probability estimates for both known and novel variants.

**Basis:** We have previously demonstrated the automated interpretation of genetic variants. We will now complete the same process for all possible combinations in an automated process.

**Research Plan:** deepInferR will incorporate a range of specialised tools and databases (VariantAnnotation, gnomAD, ClinVar, UniProt, brms, and rstan) to create a cohesive Bayesian framework. The project will follow a structured methodology involving variant data collection and annotation, probability estimation based on population frequencies and random variant estimation, incorporation of prior biological information, and Bayesian inference to update disease probabilities. Validation and simulation processes will be conducted to ensure the accuracy and clinical relevance of the predictions. This allows us to: - 1. Pre-calculate variants for diagnosing and treating genetic diseases before observing them. - 2. Provides confidence intervals to quantify previous unknowns - an important missing feature for precision medicine.

## 2.4 [Manuscript - review] Statistical developments in variant set association testing

**Objective:** This manuscript aimed to review 20 years of statistical advancements in variant set association testing (VSAT), a critical methodology for detecting associations between genetic variants and complex traits or diseases. The goal was to provide an overview of the different types of tests and their applications, emphasizing innovations and challenges in the field.

**Basis:** Variant set association testing had evolved significantly over the years. The foundation lay in understanding the genetic architecture of complex traits and rare genetic variants, with methods developed to handle aggregated data from multiple genetic variants. This review covered burden tests, adaptive burden tests, variance component tests, and combination tests, among others, highlighting their statistical properties and practical applications.

**Research Plan:** The manuscript was structured as follows: 1. **Introduction:** Provided an overview of genetic association studies and the importance of VSAT. 2. **Burden Tests:** Examined methods like CAST and CMC that aggregate rare variants. 3. **Adaptive Burden Tests:** Discussed tests that adaptively weight variants based on their frequencies and effects. 4. **Variance Component Tests:** Reviewed tests like C-alpha and SKAT that evaluate the variance of variant effects. 5. **Combination Tests:** Analysed methods combining burden and variance component tests, such as SKAT-O and MiST. 6. **Annotation:** Explored the role of variant annotation in enhancing VSAT, using databases like gnomAD and ClinVar. 7. **GLM and Score Statistic:** Applied generalised linear models and score statistics in VSAT. 8. **Advanced Methods:** Introduced recent methods like the EC Test and ACAT, which offer improved power and flexibility. 9. **Applications and Case Studies:** Provided real-world applications of VSAT in genomic studies. 10. **Conclusion:** Summarised key findings and future directions for VSAT research.

This review synthesised existing knowledge and identified gaps, aiming to guide future research and application in genomic data analysis.

## 2.5 [Manuscript - finished] Rare variants in infection response protein pathway associated with sepsis in children.

**Objective:** This manuscript aimed to investigate the role of rare genetic variants in protein pathways associated with sepsis susceptibility in children. By focusing on infection response proteins, the study sought to uncover genetic determinants that contribute to the severity and outcomes of sepsis in pediatric patients.

**Basis:** Sepsis remains a significant cause of morbidity and mortality in children, driven by complex and individualised immune responses. Previous research has largely concentrated on common genetic variants, leaving a gap in understanding the impact of rare variants. This study uses whole exome sequencing data and advanced analytical tools to explore these rare variants, employing our new methods such as ProteoMCLustR for protein pathway clustering, SkatRbrain for statistical analysis, and ACMGuru for clinical genetics interpretation.

**Research Plan:** The research was conducted in several structured stages: 1. **Analysis Overview:** Provided a comprehensive examination of the cohort and methodology. 2. **Single-case Analysis:** Detailed individual case studies to highlight specific findings. 3. **Single Variant and gene-level analysis:** Investigated the impact of individual rare variants and their respective genes. 4. **Protein pathway construction:** Built and analysed protein pathways using novel clustering techniques. 5. **VSAT analysis:** Applied variant set association tests to identify significant genetic associations. 6. **Interpretation:** Interpreted findings within multiple contexts including clinical genetics, functional relevance, and pathway functionality. 7. **Validation:** Validation protocols using best available evidence.

The study concluded by identifying a crucial protein pathway involved in sepsis, comprising genes integral to immune regulation and response. This work underscores the importance of rare genetic variants in pediatric sepsis and provides a foundation for future personalised medical approaches.

## 2.6 [Manuscript - finished] Genome-wide association study of pediatric sepsis.

**Objective:** This manuscript aimed to identify genetic determinants associated with pediatric sepsis by performing a genome-wide association study (GWAS) on a cohort of children with culture-proven bacterial sepsis. The study sought to discover loci associated with sepsis susceptibility and disease characteristics, providing insights into the genetic basis of sepsis in children.

**Basis:** This study uses samples from the Swiss Pediatric Sepsis Study, a national multicenter cohort, to explore these genetic associations through a comprehensive GWAS approach.

**Research Plan:** The research was structured as follows: 1. **Sample collection:** Collected samples and clinical data from 650 children with sepsis and 1395 controls. 2. **Genotype quality control and imputation:** Ensured high-quality genotyping data and performed imputation using the 1000 Genomes Project reference panel. 3. **Association testing:** Conducted case-control analysis to identify loci associated with sepsis susceptibility and case-only analysis for specific disease characteristics. 4. **Results interpretation:** Identified significant associations, particularly a locus on chromosome 9 encompassing the CTNNA1 and ELP1 genes, which modulate sepsis susceptibility. 5. **Discussion and conclusion:** Discussed the implications of the findings, highlighting the genetic modulation of sepsis susceptibility and the potential for targeted interventions.

This study identified a genomic region significantly associated with pediatric sepsis, offering new avenues for understanding the genetic basis of sepsis and improving patient outcomes.

## 2.7 [Manuscript - published] You AI n't using it right - artificial intelligence progress in allergy

**Objective:** This manuscript aimed to address the limitations and potential of artificial intelligence (AI) tools like ChatGPT in the field of allergy research. By exploring a practical example involving the safety of cefazolin for patients with penicillin allergy, the study demonstrated how to effectively utilise AI for complex information retrieval and analysis.

**Basis:** The basis for this manuscript stemmed from a discussion on the limitations highlighted by Dages et al. regarding AI's performance in providing accurate medical information. This work provided a nuanced view of AI capabilities and the importance of proper query formulation.

**Research Plan:** The research was conducted as follows: 1. **Initial query and AI response:** Explored the initial query posed by Dages et al. regarding cefazolin and penicillin allergy and identified the shortcomings in the AI-generated response. 2. **Complex query formulation:** Demonstrated how to formulate a more detailed and structured query for AI to perform PubMed searches, process information, and interpret results. 3. **Automated PubMed query:** Used ChatGPT to write an R script for querying PubMed, retrieving relevant abstracts, and performing text analysis. 4. **Data processing and analysis:** Implemented term frequency and correlation analysis on the retrieved texts, visualizing results through network plots and heatmaps. 5. **AI interpretation:** Summarised the processed data using AI to generate an informed response to the initial medical query. 6. **Results and visualization:** Presented the findings, including term frequencies, co-occurrence analysis, and a final AI-generated summary. 7. **Discussion:** Discussed the practical implications of using AI for medical research, acknowledging current limitations and potential for future improvements in user-friendly tools and interfaces.

The study concluded that AI tools like ChatGPT can significantly aid in medical research and information retrieval when used with appropriate methodologies, highlighting the need for ongoing development in AI technology to enhance its accuracy and usability.

## 2.8 [Grant application] Heidi Ras

**Objective:** Develop and deploy an automated software pipeline to enhance the speed and accuracy of genomic data analysis for diagnosing and treating genetic diseases.

**Basis:** Integrating advanced genomic technologies is crucial for improving diagnostics and treatment strategies for rare genetic diseases. Current fragmented methodologies delay critical responses. GenomeSwift aims to streamline and automate these processes, providing rapid, precise insights into genetic disorders, thereby improving healthcare outcomes.

**Research Plan:** The Heidi Ras application is to develop GenomeSwift. It is planned to integrate tools like ProteoMCLustR, SkatRbrain, Archipelago, ACMGuru, and DeepInferR into a unified platform for optimised genomic data processing. Robust statistical methodologies and validation processes will ensure reliability. The project aims to automate genome analysis and integrate various tools into a single pipeline, expediting data processing and enhancing diagnostic accuracy for rare genetic diseases. A detailed timetable using CPM and PERT methods was made to ensure project completion.

## 2.9 [Grant application] NCCR assist

**Objective:** Develop a national learning health system for sepsis to improve translation of scientific discoveries into clinical practice, reducing sepsis burden and enhancing personalised management.

**Basis:** Sepsis research in Switzerland is fragmented. Integrated research is needed to translate innovations into precision medicine.

**Research Plan:** Organised into six work packages: 1. Advance early detection through biomarker and biosensor development. 2. Profile host responses to pathogens for precise disease understanding. 3. Conduct clinical trials for new diagnostic and management strategies. 4. Implement and translate findings into a learning health system. 5. Develop strategies to address long-term sepsis impacts. 6. Create smart disease surveillance systems.

The project aims to improve sepsis diagnostics and treatment, align with global health priorities, and enhance pandemic preparedness.

## 2.10 [Pipeline develop] Phase 1-2 variant analysis - ML and statistics datasets prep.

**Objective:** *This is the most critical task of our work.* This process performs comprehensive variant analysis and dataset preparation for machine learning and statistical genomics, crucial for the diagnosis and treatment of genetic diseases.

**Basis:** Our goal is to optimise the handling of the vast amount of genomic data efficiently and accurately. This project integrates advanced genomic technologies and custom-developed tools to streamline the analysis process.

**Research Plan:** 1. **Data Transfer:** Transfer DNA, RNA, and protein data from SMOC to SciCORE servers. 2. **Data Processing:** Use GATK best practices pipeline for initial data processing. 3. **Variant Annotation:** Annotate variants using Ensembl VEP with over 160 genetic and biological databases. 4. **Variant Interpretation:** Use ACMGuru, a custom-developed method, to interpret variants as disease causes. 5. **Data Output:** Standardise the output for machine learning, statistical genomics, and clinical genetics reporting.

## 2.11 [Pipeline develop] Phase 1-2 statistical genomics and clinical genetics report prep.

**Objective:** Generate detailed clinical genetics reports and analyse new statistical genomics findings tailored to Kispi's needs.

**Basis:** Accurate clinical genetics reports and robust statistical genomics analysis are critical for identifying disease-causing genetic variants and understanding their biological impact.

**Research Plan:** 1. **Clinical Genetics Reports:** Produce automated reports adhering to ACMG standards and other guidelines, identifying candidate causal genetic determinants, providing context for all evidence, and quantifying evidence reliability. 2. **Statistical Genomics Results:** Report new findings from statistical genomics projects, such as identifying common genes in specific diseases or protein pathways involved in shared mechanisms.

By following this plan, the project enhances the accuracy and efficiency of genomic data processing, leading to better diagnostic and therapeutic strategies for genetic diseases.

## 2.12 [Concept develop] SwissPedHealth contribution to data models

**Objective:** Contribute genetic data concepts to the SPHN model, ensuring comprehensive representation of omics data.

**Summary:** From June 2023 to January 2024, the SwissPedHealth project collaborated with TheHyve to develop and implement genetic data concepts into the SPHN model. This included defining minimal necessary information for tracking and analysis, emphasizing data traceability, and contributing to the FAIRification of omics data. Key features such as sequencing instrument details, SOPs, QC metrics, and metadata related to sequencing assays were incorporated. Our group extended these concepts for post-analysis reporting, enhancing the schema's capacity for detailed data representation. Workshops, meetings, and feedback sessions ensured the model's suitability and integration into the SPHN Dataset for the 2024 release.

# 3 Plans for 2025-2026

## 3.1 Future Plan: Joint analysis of SwissPdHealth phases 1-3

1. Clinical genetic reports for all patients.
2. Common core dataset for all research projects; statistics, ML, etc.
3. Publish the results of statistical genomics.

## 3.2 Future Plan: Directly integrate genetic data concepts into clinical databases

**Objective:** Integrate newly developed genetic concepts and processed variant results into clinical databases, eliminating the need for additional data engineering steps.

**Basis:** - **SwissPedHealth Contribution new data models:** From June 2023 to January 2024, we collaborated with TheHyve/SPHN to develop and implement genetic data concepts into the SPHN model, ensuring comprehensive representation of omics data. - **Pipeline Development - Phase 1-2 Statistical Genomics and Clinical Genetics Report Prep:** We generated detailed clinical genetics reports and analysed new statistical genomics findings tailored to Kispi's needs.

**Future Plan:** - **Direct Integration of Variant Results:** Use the new genetic data concepts to format variant results from clinical reports directly into structures compatible with clinical databases. - **Streamlined**

**Data Processing:** Develop automated pipelines to ensure processed variant results are automatically formatted and merged with existing clinical databases.

This plan aims to enhance the efficiency of genomic data integration, improving patient outcomes and advancing personalised medicine.

### 3.3 Future Plan: Publish our SwissPedHealth genomics system.

### 3.4 Future Plan: Publish joint genomics analysis of SwissPedHealth phase 1-3

### 3.5 Future Plan: Publish/collaborate on the joint DNA, RNA, proteomics analysis of SwissPedHealth phase 1-3 (Sean, Daphné, Vito)

### 3.6 Future Plan: Publish deepInfeR - a continuously updating database of causal inference.

This aims to overcome the most important missing feature in precision medicine today. The best methods today search for known or inferred explanation of observed genetic variants. However, the unknowns are never quantified. This results in a best possible outcome of either true positives or false positive. Our goal is to quantify the full set of measures for true positive, true negative, false positive, false negative.

This requires a Bayesian approach for pre-calculating disease probabilities of genetic variants based on evidence priors. The goal is to integrate known priors and estimated unknowns into a single framework. This results in a set of genetic evidence matrices for every variant combination and every ICD-10 code.

#### 3.6.1 Overview

##### 1. Variant Data Collection and Annotation:

- **Variant identification:** Collect all known nucleotide variants and complex variants.
- **Variant annotation:** Use tools like `VariantAnnotation` to annotate each variant with functional consequences.

##### 2. Probability estimation:

- **Frequency estimation:** Estimate the probability of each variant based on true frequency in the population using databases such as gnomAD.
- **Random variant estimation:** Estimate the probability of random, novel variants.

##### 3. Prior information incorporation:

- **Biological data integration:** Integrate biological data (clinical pathogenicity, protein structure, etc.) as priors.
- **Bayesian priors:** Use priors from databases like ClinVar for pathogenicity and structural data from UniProt.

##### 4. Bayesian inference:

- **Disease Probability Calculation:** Apply Bayesian methods to update the probability of each variant causing specific diseases.
- **Posterior Probability Calculation:** Use the Bayes theorem to combine prior knowledge and observed data.

##### 5. Tools and Packages in R:

- **VariantAnnotation:** For annotating genetic variants.
- **gnomAD:** To get population frequency of variants.
- **ClinVar:** To get clinical significance of variants.
- **UniProt:** To get protein structural information.
- **brms:** For Bayesian regression modeling.
- **rstan:** For Bayesian inference using Stan.



### **3.7 Future needs: SciCORE environment**

- I am organising this and will have cost summary ready. Storage and compute costs.

### **3.8 Future needs: Research projects**

- 1-2 PhD projects for improving research quality?
- We need to recruit more - teaching students from UZH / EPFL.

### **3.9 Future needs: Precision medicine**

- Return reports for use in clinical genetics.
- Test the turnaround time: sample -> SMOC -> pipeline report.
- Get commitments to scale.