Part B: Project description

1. Executive summary

Goal of the NDS (1 paragraph):
Infections show a range of endo- and phenotypes with variable impact on **clinical course and outcomes**. To account for heterogeneity in a data-driven manner and thereby, achieve **personalized assessment, characterization and outcome prediction**, clinical data warehouses (CDWH), and our driver project Personalized Swiss Sepsis Study (PSSS) enabled major advancements in accessing, collecting, and structuring data on host, pathogen characteristics, and temporal developments. During PSSS, it became evident that **two key data gaps remain**: a) lack of clinical context and b) missing interpretation of clinical phenotypes. Profound understanding "which intervention was done and for what reason" is critical for patient management. Therefore, we will develop **harmonized clinical annotation** for infections and outcomes, as well as for **infection-related context and interpretation**. The resulting data will enable contextual data from clinical experts and data-driven approaches for patient assessment, characterization, and outcome prediction. Our **approach will revolutionize prediction of infection-related outcomes on intensive care units (ICUs)**. **First goal: Maintain the PSSS platform and expand** beyond sepsis to other infections, ensuring **FAIR principles and collaborations**. **Second goal: Focus on data quality** by harmonizing annotation of phenotypes, context, and interpretation with shared standardized definitions and ontologies. **Third goal: Form a sustainable network between ICUs, infectious diseases, microbiology, and data science.** Improving **accessibility** of data **and tools** to promote research. Contemporaneously, forming a national and international research platform. **Fourth goal:** Develop both clinical and data-driven procedures for **rapid and precise assessment** of patients exhibiting infection-related phenotypes. **In-depth and multi-dimensional characterization** of endo- and phenotypes including clinical presentation, -omics, and pathogen properties. **Early prediction** of outcomes, accounting for individual clinical context. Validation of digital biomarkers and evaluation of **potential treatment implications** e.g., focusing on antibiotic stewardship. **Fifth goal: Generation** of a multicentric FAIR **public data repository** ready for 1st/2nd/3rd party use focusing on outcomes in ICUs, with **software package** for phenotype assessment, characterization, biomarker discovery, and retrospective clinical validation.

Executive summary:
Infections show a range of diverse phenotypes with variable impact on clinical course and outcomes. Our NDS proposal focuses on this heterogeneity within **critically ill patients with severe infections** using a combined clinical- and data-driven approach for an improved personalized assessment, characterization and outcome prediction on patients with infections. We will integrate rich and diverse data sources to predict outcomes relevant in the lighthouse project and nested projects.

Organization: Our proposal is organized **similar to our SPHN/PHRT driver project PSSS** and **utilizes** already **established organization/governance structures and ethical/legal documents** to maintain and expand a sustainable data exchange platform. **Seven work-packages** define responsibilities, milestones, and deliverables for the lighthouse and nested projects. The goals of our proposal are the (i) **continuation** of **legal, governance, and IT infrastructures** between collaborators, including local CDWH teams; (ii) **expanding** and integrating **new data types** such as bacterial sequencing and MALDI-TOF mass spectrometry (MS) spectra data via the Swiss Pathogen Surveillance Platform (www.spsp.ch; SPSP); (iii) **increasing data quality** and interoperability, by including context, interpretative information, and quality

control procedures; (iv) **assessment, characterization, and validation** of clinical and data-driven phenotypes **to predict infection-related outcomes**; and (v) **building a** multicenter **public data repository and accompanying machine learning (ML) toolbox** for tasks such as clinical outcome prediction, digital biomarker discovery and patient state representation learning.

Governance: Our PSSS/SPSP consortia are already organized with **scientific and executive boards** according to **consortia, data transfer and use, and publication agreements, and ethical documents.** The consortia hold monthly meetings to inform about progress and ensure scientific exchange. The boards discuss and vote on project ideas, legal and governance matters, and data quality. Procedures to access data, publication, and induction of new projects have been established.

Data flow and data management: We will build upon existing mechanisms from PSSS and SPSP that will be extended and interconnected. Electronic case report forms (eCRF) will be used to collect clinical data at each center, and CDWH will ensure data quality and integrity and transfer data in a resource description framework (RDF)-compliant format to the research tenant. SPSP will receive genomic and MALDI-TOF data that will be processed using harmonized pipelines and then transferred in a RDF-compliant format to the research tenant. A common identifier will enable merging the datasets coming from both sources. Metadata will be exchanged with the CDWH. Within PSSS, we **reached major milestones**: (i) **Establishment** of a Swiss-wide **ethical and legal framework**; (ii) **Definition** and centralized integration of **sepsis-related datasets**; (iii) **Exchange and management of data** with 17'000+ ICU patients and 120'000+ SARS-CoV-2 sequences on SPSP. Our NDS will further expand data flow and data management along the lighthouse and nested projects, as we will build upon existing mechanisms from PSSS and SPSP that will be extended and interconnected. A common identifier will enable merging the datasets coming from both sources. Metadata will be exchanged with the CDWH.

Aims of lighthouse research project: We will (i) **define, standardize, document, and predict infection-related endotypes, phenotypes, and outcomes** such as **sepsis, immune dysregulation, acute infection-related encephalopathy, and infection-caused mortality,** as well as contextual markers on **clinical reasoning and interpretation**; (ii) **transfer and access data** from the CDWHs to the data hosting BiomedIT node including annotated phenotypes and outcomes; (iii) **develop new approaches** to discover **digital biomarkers, and predict outcomes**; (iv) **provide feedback** to improve data quality control procedures; and (v) validate digital biomarkers and evaluate **potential treatment implications**. **Nested projects** will focus on (i) the identification of endotypes of sepsis and their on the clinical management and outcomes; (ii) refining our understanding of virulence genes and how they impact outcomes through the exploration of the impact of phenotypic resistance testing on treatment; (iii) assessment of viral and fungal infection on infection-related outcomes; and (iv) determining the effects of treatments and antibiotic stewardship, resistance, and virulence of microbes. Nested projects will benefit from the ML tools developed within the framework of the lighthouse project. They will interconnect and **enrich data quality and interoperability**, as well as increase generalizability of biomarker-discovery results. This dataset guarantees future research also via 3rd party usage.

Envisioned outcomes: Primary outcome: Sustainable national infrastructure for data exchange focusing on **critical ill patients, infections, and outcomes**. Provide local, national, and international opportunities for research and patient care evaluations. **Second outcome:** Generation of a multicentric, **high quality, curated dataset** including more than 25'000 ICU patients (5 patients per day per center for 3 years). Development of procedures for curated and systematic data collection and quality controls to reduce maintenance. **Third outcome:** Discovery of **digital biomarkers** using data-driven approaches with ML tools for assessment and characterization of infection-related phenotypes, as well as risk assessments and outcome predictions. Digital biomarkers will be validated in retrospect and could be integrated into clinical support systems, for example, to optimize antibiotic treatment of sepsis. **Fourth outcome:** Formation of a

multicentric **public data repository** for 3rd party usage with anonymized data (comparable to MIMIC-IV), as well as an accompanying ML toolbox for e.g. biomarker discovery. This will **generate a global impact** for outcome research on ICUs with high quality Swiss data. Collected ICU data could also be used to investigate other diseases, such as acute heart failure, acute kidney injury, stroke.
Sustainable impact on the research community: We will build an environment for reproducible and standardized outcome research in critically ill patients. Targeted populations will benefit from novel digital biomarkers, classification, characterization, and **personalized risk assessments**. **Local institutions/hospitals** will gain important quality feedback for diagnosis, treatment, and prognosis. This will improve patient care and treatment quality on ICUs, and ultimately reduce costs. The resulting business intelligence data may result in continuous support from hospitals for CDWHs beyond the NDS funding period. In addition, we will **generate a public data repository and open-source Machine Learning software toolbox for biomarker discovery, patient state representation learning and endpoint prediction** for national and international academic partners.

International benchmarking (1 paragraph):
Our new NDS "IICU" dataset will be generated based on previous work. This dataset will be made available for international research. Compared to similar initiatives like MIMIC or eICU, our dataset will (i) also contain rich molecular data on pathogens e.g., bacterial whole genome sequencing and MALDI-TOF MS spectra. This will notably enable investigating further aspects like the impact of antimicrobial resistance and virulence in ICU patients. (ii) In addition, domain knowledge will also be integrated to the dataset, thanks to the prospectively collected contextual information. (iii) For the established data concepts focusing e.g. on laboratory, microbiology, treatment data, regular iterative quality controls ensure a low amount of missing data, which is an issue in many other databases. Through a collaboration with the 101 Fund, an international foundation that brings together a community of survivors of resuscitation in 66 countries including Switzerland. This will integrate patient related topics into our NDS proposal.Altogether, this will set a new standard in high-quality ICU datasets and provide an excellent benchmark for developing novel machine learning models and foster biomarker discovery in an unprecedented fashion. Within the SPHN/PHRT & NDS consortia, strong emphasis has and will be put in ensuring that the produced data adhere to international standards notably in terms of ontologies (e.g. ATC, LOINC, SNOMED CT) and FAIR data. The project is built upon the "SPHN/PHRT driver project on sepsis" and the "Swiss Pathogen Surveillance Platform" (PSPS). PSPS was launched in 2018 and became in 2021 the national SARS-CoV-2 genomic data hub for Switzerland and the Federal Office of Public Health. Today, SPSP is extremely well integrated within the Swiss microbiology landscape with international data sharing interfaces to ENA and GISAID. SPSP has also proven its utility and efficiency internationally, and SIB (A. Lebrand) was therefore invited to co-lead a European Elixir consortium to support nascent and established national SARS-CoV-2 sharing platforms.