

## Executive Summary

# Technical report of the implementation and validation of the SIHOS-Database.

## Project «Social Inequalities and Hospitalisations in Switzerland»

A project of the National Research Programme "Smarter  
Health Care" (NRP 74)

Olten, 10<sup>th</sup> of may 2021

Lucy Bayer-Oglesby  
Andrea Zumbrunn  
Nicole Bachmann



University of Applied Sciences and Arts Northwestern Switzerland  
School of Social Work



**Gesundheitsversorgung**  
Nationales Forschungsprogramm



Schweizerisches Gesundheitsobservatorium  
Observatoire suisse de la santé  
Osservatorio svizzero della salute  
Swiss Health Observatory



Berner  
Fachhochschule

Swiss TPH



Swiss Tropical and Public Health Institute  
Schweizerisches Tropen- und Public Health-Institut  
Institut Tropical et de Santé Publique Suisse

Associated Institute of the University of Basel

Executive Summary of the following report:

*Bayer-Oglesby L, Zumbrunn A, Bachmann N. Implementierung und Validierung der SIHOS-Datenbank. NFP74-Studie «Social Inequalities and Hospitalisations in Switzerland» «SIHOS». Olten: HSA FHNW; 2019 Aug.*

## **1 Introduction**

Within the framework of the NRP74-project «Social Inequalities and Hospitalisations in Switzerland» (SIHOS), existing national inpatient and social data have been combined for the first time in Switzerland in anonymised form into a database to analyse associations between social situations, hospitalisations and health outcomes. The SIHOS project focusses on the provision of medical care for vulnerable groups affected by chronic conditions in inpatient settings. SIHOS is a collaborative project, carried out by the Institute for Social Work and Health, FHNW School of Social Work together with the Swiss Health Observatory, the Bern University of Applied Sciences and the Swiss Tropical and Public Health Institute. This report provides an in-depth documentation of the implementation and the validation of the SIHOS-Database.

## **2 Linkage of data in anonymised form**

The linkage of national health and social data, collected by the Federal Statistical Office (FSO) for different statistics, had been a big challenge for the implementation of the SIHOS-Project. Consultations for legal issues as well as technical problems resulted in a delay of about one year. After signing the data protection contract, the necessary data sets were finally anonymized by the FSO with hashing procedures and delivered to the project team.

For creating the SIHOS-database, about 1.2 million records of the structural survey (SE 2010-2014) were linked by the anonymous linkage code with 9.6 million records of the medical statistics (MS 2010-2016), 1.0 million records of the statistics of socio-medical institutions (SOMED 2010-2016), 0.4 million dates of death of the vital statistic (BEVNAT 2011-2016) and 1.0 million dates of moving out of Switzerland of the population and household's statistics (STATPOP-Movement 2011-2016). Each year of the SE (2010, 2011, 2012, 2013 and 2014) was linked with all MS/SOMED records 2010-2016 and an indicator for matched records was created. For SE-participants with several hospitalisations, each MS- and SOMED record was linked with the corresponding SE-information. The five linked SE-files are the basis for selecting and combining records of interest for different perspectives of analysis. The file of the inpatient cohort contains only the matched records of social and inpatient data (N=987'552 records) while the file of the population cohort contains all SE-records 2010-2014 with indicators for hospitalisations in the two years following SE-participation (N=1.2 million records).

## **3 Validation of completeness and correctness of linkages**

During the approval- and data linkage-process it became apparent, that only a linkage rate of about 70 percent could be reached between the anonymized social and medical records. The question arose, whether these 70 percent of successfully linked records (Matched Records) were selected at random – and thus all participants of the SE-surveys 2010-2014 with hospital stays had the same chance to get into the SIHOS-sample – or if the selection of the 70 percent were subject to systematic bias, resulting in an under- or over-representation of certain groups in the SIHOS database. Therefore, the SIHOS-database underwent comprehensive validation and the matched records were evaluated for completeness and correctness in order to assess the potential and limits of the SIHOS database for the planned analysis.

The report evaluates causes and extent of the following two issues:

Missing Matches: Are there missing linkages between the social and medical records of participants of the structural survey who had a hospital stay?

Mismatches: Are there erroneous linkages between social and medical records, i.e., have two different individuals been attributed the same anonymous linkage code and thus been coded as one person?

Please note that the present evaluation refers to the record-level. The consistency of multiple stays of patients will be evaluated in the framework of the planned longitudinal analysis.

### 3.1 *Missing Matches*

Hospitalisation rates of SE-participants that are calculated with the SIHOS-database are about 35 percent lower compared to real rates observed in the Swiss populations 15plus. The following three factors were identified to contribute to this discrepancy: (1) erroneously built anonymous linkage codes, (2) underrepresentation of persons with severe health problems in the SE-survey (3) demographic deviations of the SE-sample from the reference population due to non-response and increase of samples in certain cantons.

The extent of erroneous linkage codes could be evaluated based on deaths that occurred in Swiss hospitals and nursing homes, since these are registered in two full surveys: on the one hand in the MS/SOMED and on the other hand in the BEVNAT statistics. For 70 percent of those who had died according to the full MS- and SOMED- surveys, the date of death of the BEVNAT could be assigned. This implies, that 30 percent points of the lower hospitalisation rates result from erroneous linkage codes. Distributions of matching rates show that the problem of erroneous linkage codes probably lies on two levels: on the one hand on the hospital level, where certain hospitals have (very) low matching rates in certain years, and on the other hand on the individual level, when not very common names, that are more frequent for Non-Europeans, are spelt in different ways. However, multivariate analysis suggest that the erroneous linkage codes most likely are randomly distributed regarding the social factors of interest, with the exception of some specific migration groups with Non-European countries of origin, showing lower chances of matching, and being therefore underrepresented in the SIHOS-sample. The planned analysis on *relative risks* by social status and for most social groups however should not be impaired in both, the population and inpatient cohort.

The underrepresentation of persons with health problems in the SE-sample can be explained by the exclusion of permanent inhabitants of socio-medical institutions from the structural survey. Since this population is not subject of the SIHOS-study, the limitation of the sample to persons in private households does not pose a relevant problem. The demographic bias of the SE-sample compared to the reference population can be explained by non-response and increased samples in certain cantons. This resulted – similar to other survey-based studies and independent of the matching process - in an over- or underrepresentation of certain population groups. Again, the planned analysis on *relative risks* by social status and for different social groups should not be impaired by this.

In the SIHOS-sample of matched records, for more than 95 percent of those persons who had died according to the MS and SOMED respectively, the corresponding date of death of the BEVNAT could be assigned. Since these records had already been successfully linked with the SE, individuals of the SIHOS patient cohort most probably have a valid anonymous linkage code. Thus, mortality analysis as well as taking account of censoring in the planned longitudinal analysis should both not be impaired.

### 3.2 Mismatches

There was a very high agreement between demographic variables that were available in both, the SE-survey and the MS (sex, age and nationality) and SOMED (sex and age) respectively. Only 0.2 percent of MS-records and 0.01 percent of SOMED-records had to be excluded because of disagreement of age and/or sex. This low rate of incorrect matches corresponds to the conservative approach of the hashing procedure. It was less clear if disagreements of nationality between SE and MS (0.7%) was also due to wrong matching, since about 17 percent of the Swiss population has a dual nationality. Therefore, MS records with plausible data on age and sex, but discrepancies regarding nationality, were kept in the SIHOS Basis Sample, yet marked, and can be excluded for sensitivity analysis (SIHOS Sensitivity Sample). Overall, the extent of mismatches can be neglected and does not at all impair findings of the SIHOS project.

## 4 SIHOS-reference file

The SIHOS reference-file of the final SIHOS-Basis-sample includes a total of N=987'552 records of N=413'589 persons. The medical statistics contributes N=950'182 records, the statistics of socio-medical institutions N=37'370 Records. The reference file includes all variables regarding demographic and social factors, information on hospital stays and stays in nursing homes as well as indicators for death and moving abroad. Further, derived variables are available regarding sequence of stays, rehospitalisations, on NCD and CCS diagnosis etc. A variable list with the available categories, the distributions in the SIHOS-Basis-Sample and missing values is presented in the appendix of this report.

## 5 Conclusions

The implementation of the SIHOS-database resulted in a unique basis to investigate associations between social situations, hospitalisations and health outcomes in inpatient settings. The size of the SIHOS Basis sample should enable analysis with specific chronic conditions and rather small vulnerable groups.

The validation of the SIHOS database for completeness and correctness of linkages showed that most analysis should be possible as planned. In analysis with the population cohort though, hospitalization rates will be underestimated by about one third, given the missing matches and thus cannot be calculated, but possibly extrapolated. In publications, the limitations of such estimates need to be indicated. However, we found no evidence that the erroneous linkage codes lead to a bias regarding the social factors of interest in the SIHOS database, with the exception of an underrepresentation of Non-European migration groups. Therefore, we consider analysis with the population cohort of *relative hospitalisation risks* by social status and for most vulnerable groups feasible.

Because of the matching rate of only 70 percent, the inpatient cohort is smaller than could be expected given the sample size of the SE and the hospitalisation rates observed in the Swiss population, and as in many surveys that are based on random-sampling, certain population groups may be under- or overrepresented. The planned analysis with the inpatient cohort on utilization and outcomes of hospitalisations though, should hardly be impaired. Since the inpatient cohort (i.e. the SIHOS-Basis-sample) consists of individuals with a valid anonymous linkage code, in particular mortality analysis as well as controlling for censoring in longitudinal analysis should be possible as planned.

Finally, the validation of the correctness of the linkages has shown that the extent of mismatches is marginal and does not at all impair the validity of analysis with the SIHOS-database.