CAS AML, project

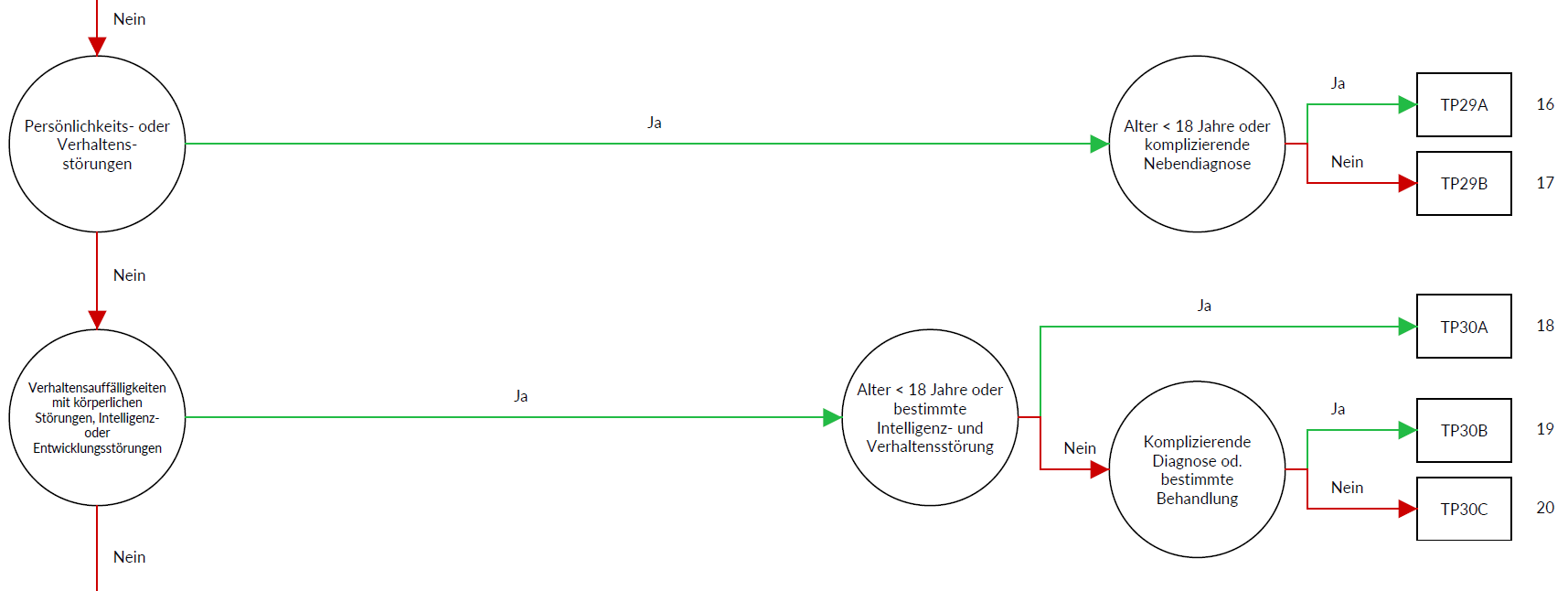
# Introduction

In an annual cycle, SwissDRG AG collects patient data from all hospitals in Switzerland in the area of fully inpatient acute somatic, psychiatric and rehabilitation care. The patient data includes

* general patient information
* medical diagnosis and service data
* cost data at case level.

With the help of this data, the tariff structures valid throughout Switzerland are developed, which are used to compensate the costs incurred in these fully inpatient areas. The TARPSY tariff structure regulates these payments for the psychiatric sector.

TARPSY divides the patients accruing throughout Switzerland into patient groups (Psychiatric Cost Groups, PCG) that are as homogeneous as possible on the basis of medical characteristics. This medical grouping logic can be represented in the form of a decision tree and is further developed annually by medically experienced specialists. Figure 1 shows a section of the grouping logic with PCGs TP29A through TP30C. Homogeneous patient groups differ on the one hand by their medical information (business know-how), and on the other hand by different average daily costs.

Figure 1: Extract from the grouping logic TARPSY, version T4.0 [1].

## ICD

An important role in the classification of patients is played by diagnoses, which are recorded according to the International Statistical Classification of Diseases and Related Health Problems (ICD) [2]. The ICD catalog describes all existing medical diagnoses and includes over 13,000 alphanumeric codes. Any number of ICD codes can be coded per case, but at least one code (principal diagnosis).

## CHOP

## In addition to the diagnosis codes, there are the service codes, which are recorded according to the Swiss surgical classification (CHOP) [3]. The same principle applies here as for the ICD codes: all services provided are recorded at case level. Thus, there can be a list of any number of CHOP codes per case. There are over 14,000 alphanumeric CHOP codes.

# Goal of the project

In this project, an algorithm called "Smart Split Detector" (SSD) is to be developed, which can be executed at any point in the decision tree and which is to provide data-based suggestions for new splits/split extensions at this very point.

The output of the SSD can be a list consisting of codes or other patient characteristics, which in the application lead to the new split better differentiating the data. By better differentiated I mean

* In the case of a differentiation of an existing split condition: A reduction of the local MSE of the new children compared to the old children.
* in the case of a new decision tree node: a reduction of the local MSE of the new children compared to the parent.

Challenges

A dummy approach would be to brute force test all patient information. This approach would make sense if there was one piece of information per patient and feature. However, since the most interesting features, ICD and CHOP codes, are in the form of lists, this approach does not seem to be the most effective because:

* Many of these codes are rarely (to never) recorded
* Sparse data can lead to statistical uncertainty when modeling cost relationships

It will therefore be necessary to develop smart lists of codes (SL), which are human-readable (do not contain too many codes) and can still detect cost relationships in the data (Smart List Detector, SLD).

## SLD

A SL contains a certain number of codes (ICD or CHOP) for which certain cost correlations have been detected in the data and which can therefore be promising for a new split/split extension.

The main goal of the project will be to use ML to generate such SLs, which can then be used as input for the SSD.

# References

[1] <https://www.swissdrg.org/application/files/3316/2999/8618/TARPSY_4.0_PV2_Definitionshandbuch.pdf>, 12.11.2021

[2] <https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-gm/kode-suche/htmlgm2021/>, 12.11.2021

[3] <https://www.bfs.admin.ch/bfs/de/home/aktuell/neue-veroeffentlichungen.assetdetail.18304268.html>, 12.11.2021

CAS AML, Projektarbeit

# Einführung

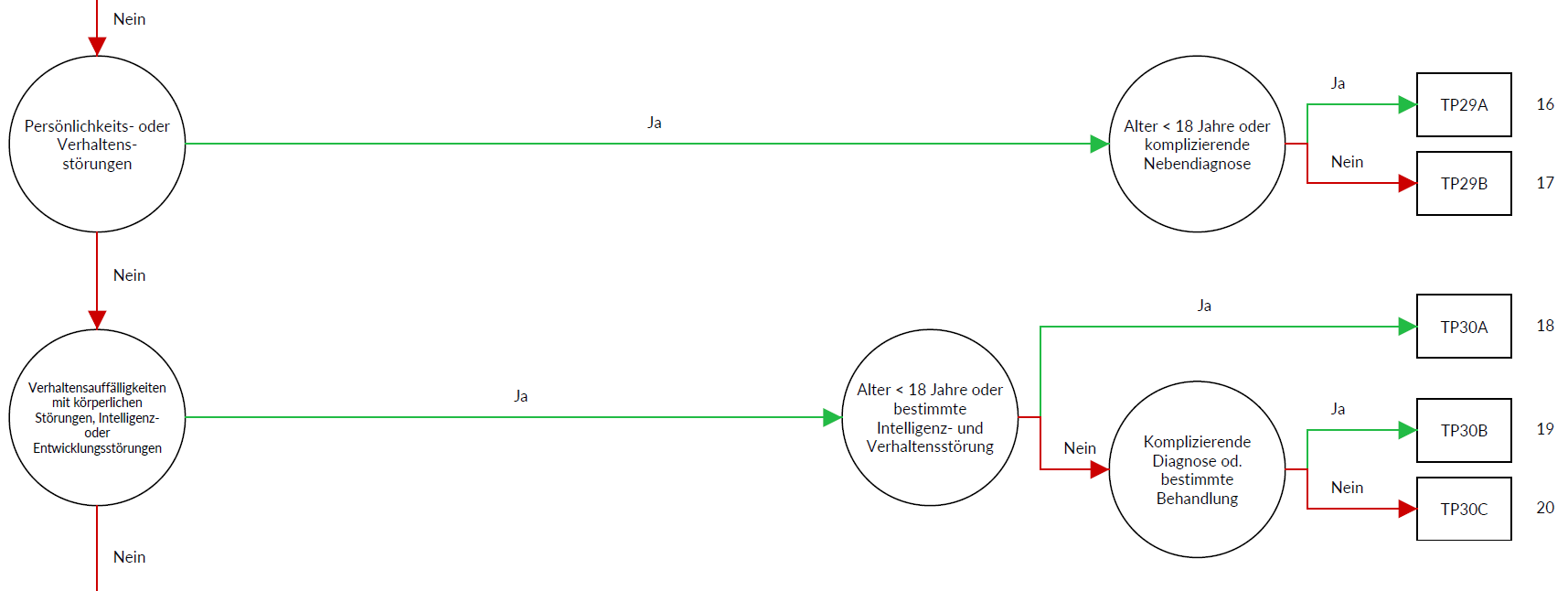
Die SwissDRG AG erhebt im Jahreszyklus die Patientendaten sämtlicher Spitäler in der Schweiz im Bereich der vollstationären Akutsomatik, Psychiatrie und Rehabilitation. Die Patientendaten beinhalten

* allgemeine Patienteninformationen
* medizinische Diagnose- und Leistungsdaten
* Kostendaten auf Fallebene.

Mithilfe dieser Daten werden die schweizweit gültigen Tarifstrukturen entwickelt, welche zur Abgeltung der angefallenen Kosten in diesen vollstationären Bereichen zur Anwendung kommen. Die Tarifstruktur TARPSY regelt diese Vergütungen für den Bereich der Psychiatrie.

TARPSY unterteilt die schweizweit anfallenden Patienten aufgrund von medizinischen Merkmalen in möglichst homogene Patientengruppen (Psychiatric Cost Groups, PCG). Diese medizinische Gruppierungslogik kann in der Form eines Entscheidungsbaumes dargestellt werden und wird von medizinisch versiertem Fachpersonal jährlich weiterentwickelt. Grafik 1 zeigt einen Ausschnitt der Gruppierungslogik mit den PCGs TP29A bis TP30C. Homogene Patientengruppen unterscheiden sich zum einen durch ihre medizinischen Informationen (Business Know-How), sowie durch unterschiedliche durchschnittliche Tageskosten.

Grafik 1: Ausschnitt aus der Gruppierungslogik TARPSY, Version T4.0 [1]



## ICD

Eine wichtige Rolle bei der Klassifizierung von Patienten spielen die Diagnosen, welche gemäss der internationalen statistischen Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (ICD) erfasst werden [2]. Der ICD Katalog beschreibt alle existierenden medizinischen Diagnosen und umfasst über 13'000 alphanumerische Codes. Pro Fall kann eine beliebige Anzahl ICD Codes codiert werden, mindestens aber ein Code (Hauptdiagnose).

## CHOP

Neben den Diagnosecodes gibt es die Leistungscodes, welche gemäss der Schweizerischen Operationsklassifikation (CHOP) erhoben werden [3]. Hier gilt dasselbe Prinzip wie bei den ICD Codes: Sämtliche erbrachten Leistungen werden auf Fallebene erfasst. Pro Fall kann also eine Liste an beliebig vielen CHOP Codes vorliegen. Es gibt über 14'000 alphanumerische CHOP Codes.

# Ziel der Projektarbeit

In dieser Projektarbeit soll ein Algorithmus namens «Smart Split Detector» (SSD) entwickelt werden, welcher an beliebiger Stelle im Entscheidungsbaum ausgeführt werden kann und an ebendieser Stelle datengestützte Vorschläge für neue Splits/Spliterweiterungen liefern soll.

Der Output des SSD kann eine Liste bestehend aus Kodes oder anderen Patientenmerkmalen sein, welche in der Anwendung dazu führen, dass der neue Split die Daten besser ausdifferenziert. Unter besser ausdifferenziert verstehe ich

* im Fall einer Ausdifferenzierung einer bestehenden Splitbedingung: Eine Reduktion des lokalen MSE der neuen Children im Vergleich zu den alten Children
* im Fall eines neuen Entscheidungsbaumknotens: Eine Reduktion des lokalen MSE der neuen Children im Vergleich zum Parent

## Herausforderungen

Ein Dummy-Ansatz würde darin bestehen, sämtliche Patienteninformationen Brute-Force durchzutesten. Dieses Vorgehen würde Sinn machen, wenn pro Patient und Feature eine Angabe vorliegen würde. Da aber die interessantesten Features, ICD und CHOP Codes in Form von Listen vorliegen, scheint dieser Ansatz nicht der effektivste zu sein, denn:

* Viele dieser Codes sind selten (bis nie) erfasst
* Sparse Data kann zu statistischen Unsicherheiten führen, wenn Kostenzusammenhänge modelliert werden

Es wird daher vonnöten sein, smarte Listen von Codes (SL) zu erarbeiten, welche zum einen human-readable sind (nicht zu viele Codes beinhalten) und trotzdem Kostenzusammenhänge in den Daten detektieren können (Smart List Detector, SLD)

## SLD

Eine SL beinhaltet eine bestimmte Anzahl an Codes (ICD oder CHOP), für welche bestimmte Kostenzusammenhänge in den Daten detektiert wurden und welche somit vielversprechend für einen neuen Split/Spliterweiterung sein können.

Hauptziel des Projektes wird darin bestehen, mittels ML solche SL zu generieren, welche dann als Input für den SSD verwendet werden können.