

Vogel, Justus; Cordier, Johannes

## Working Paper

# Application of positive and unlabeled learning: A novel approach for identifying sepsis cases from hospital administrative data

Working Paper Series in Health Economics, Management and Policy, No. 2024-02

## Provided in Cooperation with:

University of St.Gallen, School of Medicine, Chair of Health Economics, Policy and Management

*Suggested Citation:* Vogel, Justus; Cordier, Johannes (2024) : Application of positive and unlabeled learning: A novel approach for identifying sepsis cases from hospital administrative data, Working Paper Series in Health Economics, Management and Policy, No. 2024-02, University of St.Gallen, School of Medicine, Chair of Health Economics, Policy and Management, St.Gallen

This Version is available at:

<https://hdl.handle.net/10419/300110>

### Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

### Terms of use:

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



<https://creativecommons.org/licenses/by-nc-nd/4.0/>

**Working Paper Series in Health Economics, Policy, and Management**

**2024 – Nr. 02**

**Application of positive and unlabeled learning: A novel approach  
for identifying sepsis cases from hospital administrative data**

**Justus Vogel, Johannes Cordier**



# Working Paper Series in Health Economics, Policy, and Management

## Editor

Prof. Dr. Alexander Geissler  
Professor  
Chair of Health Economics, Policy, and Management  
School of Medicine  
University of St.Gallen

## Editorial office

Jonas Subelack  
Research Assistant  
Chair of Health Economics, Policy, and Management  
School of Medicine  
University of St.Gallen

The entire series of publications is available on our website at: <https://med.unisg.ch/en/research/health-care-management/publications/>

# **Application of positive and unlabeled learning: A novel approach for identifying sepsis cases from hospital administrative data**

**Keywords:** Positive and Unlabeled Data, Weakly Supervised Learning, Hospital Administrative Data, Sepsis

**JEL Classification:** I10, I19

## **Authors:**

Justus Vogel

PostDoc/ Scientific Project Leader

Chair of Health Economics, Policy, and Management, School of Medicine, University of St.Gallen

justus.vogel@unisg.ch

Johannes Cordier

Research Assistant

Chair of Health Economics, Policy, and Management, School of Medicine, University of St.Gallen

johannes.cordier@unisg.ch

## **Recommended citation:**

Vogel, Justus; Cordier, Johannes (2024): Application of positive and unlabeled learning: A novel approach for identifying sepsis cases from hospital administrative data, Working Paper Series in Health Economics, Management and Policy, No. 2024-02, University of St.Gallen, School of Medicine, Chair of Health Economics, Policy and Management, St.Gallen

# Application of positive and unlabeled learning: A novel approach for identifying sepsis cases from hospital administrative data

Dr. Justus Vogel<sup>\*c</sup> and Johannes Cordier<sup>\*</sup>

<sup>\*</sup> Chair of Health Economics, Policy and Management, School of Medicine, University of St. Gallen, St.-Jakob-Strasse 21, CH-9000 St. Gallen, Switzerland

<sup>c</sup> Corresponding author

## Declarations

## Acknowledgements

We thank the Swiss Federal Statistics Office for providing the dataset “Hospital Case Cost Statistic (*Fallkostenstatistik*)” for data years 2017 to 2019.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Conflict of interest

None.

## Author contributions

JV: Conceptualization, Methodology, Formal analysis, Interpretation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project administration

JC: Conceptualization, Methodology, Formal analysis, Data curation, Interpretation, Writing – Review & Editing

**Ethics approval and consent to participate:** Not applicable.

**Consent for publication:** Not applicable.

**Data availability:** The data is available from the Swiss Federal Statistics Office upon request.

**Code availability:** The code will be uploaded to GitHub upon project completion.

## Abstract

Only positive instances of various events, e.g., secondary diagnoses, are actively labeled in hospital administrative data. In line with this, several studies indicate underreporting of adverse events such as sepsis. The gold standard for relabeling of uncoded sepsis cases, medical record review, is laborious, costly, and infeasible to execute for identifying sepsis in large, national datasets. We apply a positive unlabeled (PU) learner as a novel approach to identify sepsis cases from hospital administrative data.

We exploit the Hospital Case Cost Statistic from the Swiss Federal Statistics Office (data years 2017 to 2019) including 72 cost attributes at case level. We hypothesize that these cost data should prove effective for learning a classification model as positive sepsis cases in the unlabeled data should exhibit similar cost patterns as labeled positive examples. We randomly draw 200,000 unlabeled examples from the full dataset and add 64,915 positive examples of sepsis labeled in the observation period for model training and evaluation. We train a robust PU learner proven in other applications, AdaSampling, with support vector machine as classification model. For model evaluation, we perform five-fold cross-validation. Due to the PU setting, we can only use positive examples in the test set and estimate recall along with precision and recall at 10%, 20%, and 30% for four different evaluation scenarios, changing the coding strategy for labeling sepsis cases.

Our model has a recall of 85.1% when labeling sepsis cases explicitly in the test set. Recall decreases to 55.5% when labeling sepsis cases exclusively with an implicit coding strategy. Recall at k% is highest for the evaluation scenarios focusing on implicit coding strategies, yet remains relatively low throughout all scenarios. Precision at k% is highest when only considering cases as positive examples that would be labeled according to both the explicit as well as implicit coding strategy (e.g., 92.3% for k=10%).

Compared to the sensitivity of directly identifying sepsis cases from hospital administrative data reported in studies using medical record review, the recall of our model is high. We propose a two step process using PU learning for increasing the quality of hospital administrative data and performing sensitivity analyses for health economic and health services research.

## 1 Introduction

Commonly, a classification algorithm that is to distinguish between positive and negative examples in unseen data, commonly referred to as binary classification, is learned from a fully labeled training set. In practice, situations arise, however, where only one class, usually positive examples, is labeled and the unlabeled data contains both negative and positive examples. Accordingly, the subject of learning from positive and unlabeled (PU) data has increasingly received attention in the machine learning research community in the last two decades (Bekker and Davis, 2020; Calvo et al., 2007a; Denis et al., 2005; Elkan and Noto, 2008; Zhang et al., 2019).

Large datasets used for health economic and health services research, e.g., from administrative and billing data, are prone to limitations regarding the documentation as well as correct and consistent coding of certain (adverse) events (Freitas et al., 2014). In fact, only positive instances are actively coded (e.g., secondary diagnoses) and uncoded instances are assumed to be negative rather than missing or unlabeled. Accordingly, negative examples in hospital administrative data might contain a considerable degree of class label noise. Consequentially, values of quality indicators used in health economic and health services research, for instance inpatient complication rates or patient safety indicators, might be higher or lower than those that are currently estimated from administrative data (Maass et al., 2015). Similarly, calculation of cost weights for and grouping of diagnosis-related groups (DRGs) could be affected by under- and misreporting. DRGs are used for hospital payment in many countries, e.g., in Switzerland and Germany, and are grouped, among other information, using diagnoses and procedure codes (Blümel et al., 2020; De Pietro et al., 2015). As both might suffer from class label noise, DRG grouping might be distorted, possibly affecting hospitals' revenues - and ultimately hospitals' profits. Sepsis is a well-researched example of the above described limitations. Different studies have found that sepsis cases are under- and/ or misreported in hospital administrative data (Fleischmann-Struzek et al., 2018; Fleischmann-Struzek and Rudd, 2023; Mellhammar et al., 2023; Rhee et al., 2015; Schwarzkopf et al., 2023). Studies often rely on medical record review by experts checking whether a non-reported diagnosis actually means that the patient did not suffer from that disease or complication. Following this approach, Mellhammar et al. (2023) find, for instance, that inpatient sepsis rates estimated from administrative data were 1.0% (95% CI, 1.0%-1.1%) to 1.4% (95% CI, 1.4%-1.5%) while medical record review estimated a sepsis rate of 4.1% (95% CI, 3.6%-4.5%). Similarly, Schwarzkopf et al. (2023) report low sensitivity (26.8% to 38.0% depending on sepsis definition) for sepsis detection via diagnosis codes from hospital administrative data. Using medical record review, the authors found that in their sample of more than 10,000 inpatient cases from ten German hospitals, incidence of severe sepsis-1 was at 3.3% instead of 1.4% as estimated from hospital administrative data.

PU learning algorithms have been applied in biology and bioinformatics, e.g., for gene prediction and detection (Calvo et al., 2007b; Wang et al., 2006) and gene network inference (Cerulo et al., 2010), in text classification (Liu et al., 2003; Yu et al., 2003), and various other applications (Jaskie and Spanias, 2019). While there are some PU learning applications investigating hospital management topics (Arjannikov and Tzanetakis, 2021a, 2021b), to the best of our knowledge, PU learning has so far not been applied to increase the quality of hospital administrative data, e.g., to improve health economic and health services research.

If discussed at all, health economic and health services studies using hospital administrative data list the issue of unlabeled data as one of their limitations. At the same time, studies cannot use laborious, costly methods such as medical record review to correctly label tens of thousands of observations or even more, e.g., in a national database. PU learning models, on the other hand, might constitute a more efficient approach to increase administrative data quality. Thus, we investigate the following research question: Is PU learning a suitable approach to increase hospital administrative data quality?

We focus on sepsis as application example due to three reasons. Firstly, it is well researched that sepsis is commonly underreported when only relying on hospital administrative data. While there is evidence that the reporting of sepsis in hospital administrative and claims data has increased in the last two decades (e.g., Fleischmann-Struzek et al., 2018; Rhee et al., 2014), the above cited, very recent studies indicate that the issue persists (Mellhammar et al., 2023; Schwarzkopf et al., 2023). Secondly, sepsis is a major complication for both surgical and non-surgical patients (Singer et al., 2016). This underscores its medical relevance and is advantageous regarding data volume PU data learners can learn from. Thirdly, sepsis has highly relevant medical and economic consequences: Patients' health status deteriorates considerably, sepsis is strongly linked to high hospital mortality rates, and the economic burden of sepsis is considerable (Rhee et al., 2014; Singer et al., 2016; Tiru et al., 2015).

## 2 Methods

### Data

#### *Learning attributes and data source*

We hypothesize that a classifier might learn positive examples of sepsis cases from hospital cost data. Costs accrue from treatment. While sepsis might not have been documented by physicians and/ or nurses and/ or it might not have been coded from documented information by coding specialists (or in fact it might not have been diagnosed at all) (Rhee et al., 2014; Schwarzkopf et al., 2023), we assume that symptoms are observed and treated. While a non-treatment might occur in single instances, it should not occur systematically. Indeed, a deliberate non-treatment of symptoms would be highly unethical. Consequentially, positive examples of sepsis part of the unlabeled data should exhibit similar cost patterns as actual positive examples.



We exploit the Hospital Case Cost Statistic from the Swiss Federal Statistics Office for data years 2017 to 2019. This dataset should be well-suited for PU data classifiers to learn from hospital cost data as it contains 71 distinct cost attributes and additionally total costs at case level. There are eleven variable, direct cost categories including, for instance, pharmaceuticals, blood and blood products, medical materials, and implants. The remaining 60 cost attributes refer to indirect costs. More specifically, in the dataset, 30 service centers are listed, comprising medical services such as intensive care unit (ICU) physician services, pathology, physiotherapy, dialysis, laboratory analyses, anesthesiology, operating room area, emergency services, and diagnostic imaging, as well as tertiary services such as patient administration and hospitality services. For each of these service centers, two indirect cost categories are given, namely one including indirect costs without asset usage costs and one including only asset usage costs.

Besides, the dataset contains information on cases' gender, age (5-year age bands), admission reason, discharge reason, main treating hospital department, diagnosis-related group code, the main diagnosis, up to 49 secondary diagnoses, the main procedure, up to 99 additional procedures, and many other patient characteristics. We deliberately do not use these attributes for learning due to two reasons: Firstly, patient characteristics such as age, gender, and co-morbidities might be well-suited to predict the risk of developing sepsis. For risk prediction, a fully labeled dataset and supervised learning methods would be used. In PU learning, the setting is distinctly different, however. In addition, further analysis with PU learner outputs might be biased if the same attributes were used in the PU learning process that would later also be used for, e.g., risk-adjustment of outcome indicators. Secondly, attributes such as the diagnosis-related group code, specific procedures, and the treating hospital department should be well-reflected in the cost attributes. For learning our PU classifier, we thus focus on cost attributes.

There are several tens of thousands of ICD-10-codes and procedure codes. Each case receives one main diagnosis and between one and 100 procedures. The number of possible combinations of main diagnosis and list of procedures is vast. It is perceivable that there are combinations that lead to similar cost patterns as seen for labeled and unlabeled sepsis cases. Thus, we complement cost attributes with aggregated information on main diagnosis and performed treatments (see Table 5 in the supplements for a full list) to enable our PU learner to efficiently learn sepsis cases from cost data.

#### *Identification of sepsis cases*

Commonly, sepsis cases are identified in hospital administrative data via diagnosis codes specified in the International Classification of Diseases, Version 10 (ICD-10). Yet there is no one gold standard what diagnosis codes or combination of codes to use for identification of sepsis cases (Singer et al., 2016). Usually, authors rely on different explicit and implicit coding strategies. Explicit coding means that ICD-10-codes concretely naming sepsis as disease are used while implicit coding strategies rely on

combinations of ICD-10-codes strongly linked to sepsis. To evaluate our PU learner, we use an explicit as well as an implicit coding strategy (Angus definition (Angus et al., 2001), i.e., organ dysfunction in combination with an infection), described in more detail below (see Table 4 in the supplements for used ICD-10-codes). Both definitions were also used by Schwarzkopf et al. (2023).

#### *Inclusion and exclusion criteria, undersampling and size of final sample*

As sepsis can potentially occur in any patient, we include all acute somatic care inpatient cases. In our dataset, age is recorded in 5-year age bands (0-4, 5-9, 10-14, 15-19, etc.). To only include adults, we exclude cases below 20 years of age. The Hospital Case Cost Statistic provided by the Swiss Federal Statistics Office includes all acute somatic care cases hospitalized in Swiss hospitals after cleaning of the cost data. In this data cleaning process, about 30% of cases are excluded.

Still, applying the above inclusion and exclusion criteria, the sample would comprise more than 2.7 million cases with an explicit (implicit) coding sepsis rate of roughly 2.39% (6.49%). To balance the dataset and manage computation time, we reduce the number of negative (or rather unlabeled) sepsis cases (undersampling). To this end, we extract all 64'915 observed positive sepsis cases (explicit coding) and add a random sub-sample of 200,000 unlabeled sepsis cases from the initial dataset. Our final sample comprises 264'915 cases with a positive class label mean of 24.50%.

#### **PU learning algorithm**

We first briefly formally review binary classification and preliminaries relevant for PU learning (e.g., the labeling mechanism). Second, we discuss the underlying assumptions for PU learning in the context of hospital administrative data. Third, we present the learning technique we applied. Lastly, we discuss model evaluation. For all four sections, we rely on the work of Bekker and Davis (2020), who have provided a thorough review of the current state of knowledge regarding PU learning.

#### *Formal descriptions*

PU learning is a type of binary classification. Yet it differs succinctly from classical set-ups of (a) fully and semi-supervised learning, and (b) learning from positive-only or one-class data. Bekker and Davis (2020) outline that main differences are that (a) only positive examples are reliable and there are no negative examples to learn from and (b) not only positive but also unlabeled examples in the training set are used for learning.

When training a binary classifier to predict the class label of an example, training examples are tuples  $(x, y)$ ,  $x = \{a_1, a_2, \dots, a_k\}$  describing a vector of attribute values of data instance  $x \in X$ , and  $y \in \{0, 1\}$  denoting the class label ( $y = 1$  indicates a positive and  $y = 0$  a negative example). To train a binary classification learner, it is assumed that the training set is an independent and identically distributed (iid) sample of the real distribution:

$$\begin{aligned} \mathbf{x} &\sim f(x) \\ &\sim \alpha f_+(x) + (1 - \alpha) f_-(x) \end{aligned} \quad (1)$$

where  $\mathbf{x}$  denotes a set of vectors of attribute values,  $\alpha = p(y = 1)$  is the class prior,  $f$  is the probability density function of the true distribution, and  $f_+$  of the positive and  $f_-$  of the negative examples.

While PU learners also aim to classify positive and negative examples subject to their attributes, they must do so from the labeled subset of positive examples and unlabeled data instances, i.e., they cannot learn from negative examples. The binary variable  $s$  indicates whether  $x$  was selected to be labeled (Elkan and Noto, 2008). In the PU data setting, the probability of an example to be selected for labeling equals 0 if it is unlabeled in the dataset:

$$p(s = 1 | x, y = 0) = 0 \quad (2)$$

Accordingly, if an example was selected for labeling, it belongs to the positive class:

$$p(y = 1 | s = 1) = 1 \quad (3)$$

A PU dataset can then be described as a set of triplets  $(x, y, s)$  with  $s$  as binary variable indicating whether the tuple  $(x, y)$  was selected to be labeled. Note also that only  $(x, s)$  are recorded and that if  $s = 0$ , an example could be positive or negative (Bekker et al., 2020; Elkan and Noto, 2008).

Bekker et al. (2020) define the propensity score, i.e., the probability for a positive example  $x$  to be selected for labeling, as  $e(x) = p(s = 1 | x, y = 1)$ . Building on the PU definition and Bayes' rule, the authors then define the labeled distribution  $f_l$  in relation to the positive distribution  $f_+$  as

$$f_l(x) = \frac{e(x)}{c} f_+(x) \quad (4)$$

with label frequency  $c = p(s = 1 | y = 1)$ , describing the fraction of positive examples selected for labeling.

### *Assumptions*

#### a. Training set

We assume a single-training-set scenario (Bekker and Davis, 2020). This means that both positive and unlabeled examples originate from the same dataset. Additionally, the dataset is an iid sample from the real distribution. Further assuming that a fraction  $c$  from the positive instances were selected for labeling according to examples' individual propensity score  $e(x)$ , the share of labeled examples in the dataset is  $\alpha e(x)$  in the single-training-set scenario.

$$\begin{aligned} \mathbf{x} &\sim f(x) \\ &\sim \alpha e(x) f_l(x) + (1 - \alpha e(x)) f_u(x) \end{aligned} \quad (5)$$

with  $f_l$  and  $f_u$  denoting the labeled and unlabeled distribution, respectively.

We assume a single-training-set scenario as positive and unlabeled examples of sepsis cases recorded in hospital administrative data all originate from the Hospital Case Cost Statistic dataset provided by the Swiss Federal Statistics Office. Data for this dataset are collected at all hospital sites in Switzerland.

Recording of positive or unlabeled examples are not limited to a subset of hospitals but each hospital can record both positive as well as negative (or rather unlabeled) instances. Single-training-set scenarios are most common in PU learning applications and thus, available learning algorithms can usually handle them well (Bekker and Davis, 2020).

b. Labeling mechanism

To enable learning from PU data, the mechanism how an example was selected to be labeled as positive must be understood for one’s dataset. As outlined earlier, we must assume a bias in diagnosis, documentation and coding of sepsis cases. First, sepsis might not have been diagnosed as, for instance, physicians did not interpret symptoms accordingly and thus did not order laboratory analyses and/ or take other diagnostic measures. Whether sepsis was diagnosed or not thus is biased by treating physicians’ experience with the disease and their behavior and decisions when observing symptoms. Second, coding relies on thorough documentation. Schwarzkopf et al. (2023) found that if sepsis was not named in patients’ medical charts, only 7.6% of true positive cases were explicitly coded as sepsis-1 cases while of the cases where sepsis was named in the medical chart, 61.8% of true positive cases were explicitly coded. Third, the findings of Schwarzkopf et al. (2023) also show that even if documented correctly, 38.2% of true positive cases were not coded, implying improvement potential in coding practices. According to Schwarzkopf et al. (2023), documentation and coding quality varied strongly between the ten hospitals participating in their study, potentially underscoring how the experience and behavior of doctors, nurses, and coding specialists influences data quality.

In PU learning, it is either assumed that positive examples were Selected Completely At Random (SCAR) or Selected At Random (SAR) (Bekker and Davis, 2020). In our dataset, the attributes potentially biasing diagnosis, documentation, and coding of sepsis cases are not recorded. For instance, there is no attribute describing physicians’ experience with sepsis cases or qualification of coding specialists translating documented information into ICD-10-codes. Still, we must follow the SAR assumption: positive examples’ probability to be labeled in our dataset depends on their attributes, i.e., our sample is a biased sample from the positive distribution (Bekker et al., 2020; Bekker and Davis, 2020):

$$e(x) = p(x, y = 1) \tag{6}$$

Bekker et al. (2020) introduced different settings in which learning under the SAR assumption with unknown exact propensity scores is possible. Specifically, in the context of our study, we may assume a reduction of SAR to SCAR as the attributes influencing the examples’ propensity score to be labeled are fewer (or in fact even different) than the attributes used for the classification model. The underlying additional assumption is that it is not possible to know if an example got labeled due to its propensity score or due to an actually low class probability. Then, there must be a subset of attributes the propensity score depends upon, i.e., the propensity attributes  $x_e$ :

$$p(s = 1 | x, y = 1) = p(s = 1 | x_e, y = 1)$$

$$e(x) = e(x_e) \tag{7}$$

In the hospital context, attributes influencing the labeling mechanism are the experience and behavior of physicians, nurses, and coding specialists (and the written and verbal communication between these groups). We may assume a reduction from SAR to SCAR, as the classification model will rely on other attributes than the labeler (i.e., the coding specialist). Specifically, the labeler does not use hospital cost data at case level as these data are irrelevant for coding.

c. Data and class distribution

Lastly, we assume separability and smoothness regarding data distribution:

- **Separability:** A classifier exists that can distinguish between negative and positive examples. Bekker and Davis (2020) formulate this assumption as a function  $f$  mapping positive examples to a value above or equal to a defined threshold  $\tau$  and vice versa for negative examples:

$$f(x_i) \geq \tau, \quad y_i = 1$$

$$f(x_i) < \tau, \quad y_i = 0$$

We expect hospital cost data to be a rich data source for training such a classifier.

- **Smoothness:** If the attributes of two instances  $x_1$  and  $x_2$  are similar, their probability to belong to the positive class will also be similar, i.e.  $p(y = 1|x_1)$  is similar to  $p(y = 1|x_2)$ . As outlined earlier, we assume cost attributes to reflect treatment of sepsis symptoms (e.g., medication, ventilation, intensive care nursing, etc.). Thus, cases with similar cost patterns should have a similar probability to belong to the positive sepsis class.

*Selected learning technique*

PU learning algorithms can be categorized into (1) one-class, (2) heuristic, (3) robust, and (4) bias-based approaches. Each of these approaches has distinct advantages and disadvantages, well-described elsewhere (Bekker and Davis, 2020; Yang et al., 2019, 2017). For our study, we rely on the AdaSampling package for the programming language R, developed by Yang et al. (2019, 2017). AdaSampling was developed building on wrapper-based feature selection (Kohavi and John, 1997). It belongs to the PU learning algorithm category of robust approaches yet it does not identify negative instances via noisy filtering or a prespecified threshold, nor does it need bias estimation like other robust approaches.

Instead, Yang et al. (2019) outline that the algorithm iteratively estimates the probability of data instances' labels to be mislabeled using a choosable learning algorithm, such as radial kernel support vector machine (SVM) or k-nearest-neighbors (k-NN). With each new iteration, the estimation is based on a resampled dataset from the initial sample with the probability of a data instance to be excluded from this resampled dataset being equal to its probability to be mislabeled. This means that data instances with a higher probability to be mislabeled are more likely to be eventually excluded from

updated training sets. The training set for the final prediction of the binary classifier then consists of examples with a relatively low probability to be mislabeled, i.e., they are reliable examples.

The AdaSampling algorithm has been used in other studies (e.g., by Zhou et al. (2022)), showing reliable results. From the set of available probabilistic classification models SVM, k-NN, feature weighted k-NN, logistic regression, and linear discriminant analysis, we use the default setting of AdaSampling, i.e., SVM for classification. We run the SVM classifier 20 times, inducing an ensemble learning model. All calculations are performed with R version 4.2.1 2021.11.01 and AdaSampling version 1.3.

### *Model evaluation*

To answer our research question, we need to assess the quality of the PU learner’s output. The main challenge is that by PU definition, datasets do not contain actual negatives. Thus, false positives (FP) and true negatives (TN) cannot be identified in PU data. Commonly, when developing PU learners, authors circumvent this evaluation issue creating synthetic datasets and/ or using benchmark datasets, e.g., when evaluating developed algorithms.

Our goal is to increase the quality of a real-world dataset, however. Hence, we must use the actual dataset in which only actual positive examples are reliable. Combined with the positive and negative predictions from our model, we can calculate recall (equal to sensitivity)  $r$ , recall at k%  $r_k$  and precision at k%  $p_k$  (Buckley and Voorhees, 2000; Manning et al., 2009), for evaluation:

$$r = \frac{TP}{TP + FN} \quad (8)$$

$$r_k = \frac{TP \text{ in top } k\%}{TP} \quad (9)$$

$$p_k = \frac{TP \text{ in top } k\%}{\text{all data instances in top } k\%} \quad (10)$$

As Bekker and Davis (2020) point out, recall can be estimated in the PU learning setting under the SCAR assumption, which is why it is essential that we can reduce SAR to SCAR as discussed earlier. We argue that for our goal, recall is the central evaluation metric as we need to correctly identify as many sepsis cases from the administrative data as possible.

We estimate  $r_k$  and  $p_k$  for the top 10%, 20%, and 30% of predictions. Both metrics were developed for information retrieval applications such as web searches. With  $r_k$ , we can evaluate what share of TPs are among the top k% of our model’s predictions. With  $p_k$ , we assess the accuracy of our top predictions.

When evaluating, we calculate the above metrics counting actual positive examples in four different constellations: (1) All cases with explicit sepsis codes, (2) cases with explicit or implicit sepsis codes, (3) all cases with implicit codes, and (4) cases with implicit codes but without explicit codes. Note that we train our model defining positive labels exclusively with explicit codes. Constellations (2) to (4) will thus

help us to evaluate whether our learner can correctly identify positive examples of an alternative sepsis coding definition in the unlabeled data.

For evaluation, we perform cross-validation with five equal folds of the positive examples. We run five iterations to train and test our model, altering with each iteration what four folds are added to the 200'000 unlabeled examples used for training and what fold is reserved for testing. We average evaluation results of the five iterations to estimate the final evaluation metrics. We use the positive class label mean of 24.50% as threshold for positive classification.

### 3 Results

#### *Descriptive results*

We present descriptive results of selected attributes in Table 1 and of the number and share per ICD-10-chapter in Table 2. We provide a full list of descriptive results in Table 6 in the supplements.

**Table 1: Descriptive results of selected attributes**

Attribute	Total sample (n=264,915)		Labeled sepsis cases (n=64,915)		Unlabeled cases (n=200,000)	
	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )
Total costs	17,895 (40,575)	8,515 (4,837- 16,610)	36,342 (71,718)	14,731 (7,720- 34,426)	11,907 (19,104)	7,418 (4,240- 13,363)
<i>Variable direct costs [CHF] – selected attributes</i>						
Pharmaceuticals	518 (3183)	89 (19-252)	1398 (5783)	261 (94- 817)	232 (1494)	63 (10-168)
Blood and blood products	227 (3055)	0 (0-0)	763 (5935)	0 (0-0)	53 (896)	0 (0-0)
Medical material	587 (2066)	77 (1-418)	1009 (3448)	108 (17- 593)	450 (1311)	66 (0-381)
<i>Fixed indirect costs [CHF] – selected attributes</i>						
Operating room, overheads excl. IUC	711 (2174)	0 (0-830)	982 (3765)	0 (0-0)	623 (1275)	0 (0-934)
Operating room, IUC	181 (516)	0 (0-197)	233 (833)	0 (0-1)	164 (355)	0 (0-225)
Operating room doctors - activities 6a, overheads excl. IUC	238 (903)	0 (0-0)	342 (1387)	0 (0-0)	204 (672)	0 (0-34)
Operating room doctors - activities 6a, IUC	13 (74)	0 (0-0)	18 (106)	0 (0-0)	11 (59)	0 (0-0)
Anesthesia, overheads excl. IUC	602 (1650)	0 (0-758)	850 (2759)	0 (0-493)	521 (1054)	0 (0-802)
Anesthesia, IUC	58 (160)	0 (0-69)	79 (263)	0 (0-39)	50 (106)	0 (0-74)
Intensive care unit, overheads excl. IUC	2165 (14418)	0 (0-0)	7592 (27602)	0 (0-3274)	403 (3925)	0 (0-0)
Intensive care unit, IUC	204 (1346)	0 (0-0)	714 (2573)	0 (0-271)	38 (374)	0 (0-0)
Intensive care unit physicians - activities 6b1, overheads excl. IUC	411 (3236)	0 (0-0)	1418 (6225)	0 (0-139)	84 (928)	0 (0-0)
Intensive care unit physicians - activities 6b1, IUC	17 (167)	0 (0-0)	56 (310)	0 (0-0)	4 (70)	0 (0-0)
IMCU, overheads excl. IUC	290 (3050)	0 (0-0)	799 (5671)	0 (0-0)	125 (1331)	0 (0-0)
Intermediate Care Units (IMCU), IUC	21 (246)	0 (0-0)	55 (455)	0 (0-0)	9 (111)	0 (0-0)

Attribute	Total sample (n=264,915)		Labeled sepsis cases (n=64,915)		Unlabeled cases (n=200,000)	
	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )
IMCU physicians - activities 6b2, overheads excl. IUC	41 (445)	0 (0-0)	107 (803)	0 (0-0)	19 (225)	0 (0-0)
IMCU physicians - activities 6b2, IUC	2 (26)	0 (0-0)	5 (46)	0 (0-0)	1 (15)	0 (0-0)
Emergency, overheads excl. IUC	196 (330)	0 (0-323)	352 (369)	307 (87-479)	145 (300)	0 (0-235)
Emergency, IUC	20 (36)	0 (0-31)	36 (41)	26 (6-53)	15 (33)	0 (0-22)
Emergency medical services - activities 6b3, overheads excl. IUC	94 (188)	0 (0-133)	167 (227)	68 (0-280)	70 (166)	0 (0-44)
Emergency med. serv. – activit. 6b3, IUC	4 (12)	0 (0-1)	7 (15)	0 (0-7)	3 (10)	0 (0-0)
Imaging procedures, overheads excl. IUC	293 (731)	55 (0-312)	657 (1179)	297 (86-744)	175 (446)	0 (0-176)
Imaging procedures, IUC	81 (212)	13 (0-82)	179 (328)	76 (23-199)	50 (142)	0 (0-46)
Laboratory, overheads excl. IUC	520 (1878)	108 (0-405)	1477 (3459)	570 (251-1270)	209 (627)	58 (0-210)
Laboratory, IUC	61 (232)	11 (0-44)	175 (428)	61 (23-147)	25 (82)	6 (0-23)
Dialysis, overheads excl. IUC	70 (1107)	0 (0-0)	239 (2137)	0 (0-0)	15 (358)	0 (0-0)
Dialysis, IUC	8 (125)	0 (0-0)	25 (238)	0 (0-0)	2 (45)	0 (0-0)
Physicians, activities 1-5, overh. excl. IUC	1411 (3168)	790 (216-1613)	2561 (5343)	1443 (738-2733)	1038 (1860)	623 (112-1289)
Physicians, activities 1-5, IUC	97 (234)	38 (2-104)	170 (360)	76 (23-175)	73 (169)	28 (0-85)
Physiotherapy, overheads excl. IUC	267 (1090)	0 (0-208)	609 (1826)	185 (0-598)	156 (665)	0 (0-114)
Physiotherapy, IUC	35 (135)	0 (0-25)	73 (190)	19 (0-74)	22 (109)	0 (0-13)
Non-medical therapies and consultations, overheads excl. IUC	148 (911)	0 (0-31)	257 (812)	3 (0-225)	112 (938)	0 (0-13)
Non-med. therapies and consultations, IUC	15 (116)	0 (0-3)	26 (97)	1 (0-20)	11 (121)	0 (0-1)
Medical and therapeutic diagnostics, overheads excl. IUC	134 (613)	0 (0-0)	234 (823)	0 (0-121)	101 (523)	0 (0-0)
Medical and therapeutic diagnostics, IUC	37 (155)	0 (0-0)	64 (212)	0 (0-31)	28 (129)	0 (0-0)
Medical and therapeutic diagnostics doctors - activities 6b5, overheads excl.	71 (412)	0 (0-0)	155 (684)	0 (0-13)	44 (265)	0 (0-0)
Medical and therapeutic diagnostics - activities 6b5, IUC	3 (22)	0 (0-0)	7 (33)	0 (0-0)	2 (16)	0 (0-0)
Nursing care, overheads excl. IUC	3705 (8127)	1582 (720-3719)	7158 (12458)	3670 (1616-8028)	2584 (5656)	1282 (630-2644)
Nursing care, IUC	256 (625)	104 (36-252)	462 (945)	210 (79-500)	189 (458)	85 (31-193)
<i>CHOP chapters: Procedures and surgeries of the...</i>						
Procedures not classified elsewhere	0.25 (0.78)	0 (0-0)	0.31 (0.94)	0 (0-0)	0.23 (0.71)	0 (0-0)
Nervous system	0.08 (0.46)	0 (0-0)	0.08 (0.58)	0 (0-0)	0.09 (0.41)	0 (0-0)
Endocrine system	0 (0.07)	0 (0-0)	0 (0.05)	0 (0-0)	0 (0.08)	0 (0-0)
Eye	0.02 (0.25)	0 (0-0)	0 (0.07)	0 (0-0)	0.02 (0.29)	0 (0-0)
Ears	0.01 (0.1)	0 (0-0)	0 (0.06)	0 (0-0)	0.01 (0.11)	0 (0-0)
Nose, mouth, throat	0.04 (0.31)	0 (0-0)	0.02 (0.24)	0 (0-0)	0.04 (0.33)	0 (0-0)



Attribute	Total sample (n=264,915)		Labeled sepsis cases (n=64,915)		Unlabeled cases (n=200,000)	
	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )
Respiratory system	0.08 (0.54)	0 (0-0)	0.22 (0.95)	0 (0-0)	0.03 (0.29)	0 (0-0)
Cardiovascular system	0.3 (1.65)	0 (0-0)	0.67 (2.82)	0 (0-0)	0.18 (0.98)	0 (0-0)
Hematopoietic and lymphatic system	0.02 (0.19)	0 (0-0)	0.04 (0.25)	0 (0-0)	0.02 (0.17)	0 (0-0)
Digestive tract	0.29 (1.25)	0 (0-0)	0.67 (2.22)	0 (0-0)	0.17 (0.65)	0 (0-0)
Urinary organs	0.11 (0.47)	0 (0-0)	0.22 (0.67)	0 (0-0)	0.07 (0.37)	0 (0-0)
Male sexual organs	0.01 (0.14)	0 (0-0)	0.01 (0.16)	0 (0-0)	0.02 (0.14)	0 (0-0)
Female sexual organs	0.04 (0.31)	0 (0-0)	0.01 (0.16)	0 (0-0)	0.05 (0.35)	0 (0-0)
Obstetric procedures	0.1 (0.48)	0 (0-0)	0 (0.08)	0 (0-0)	0.13 (0.54)	0 (0-0)
Musculoskeletal system	0.28 (1.06)	0 (0-0)	0.21 (1.23)	0 (0-0)	0.30 (1)	0 (0-0)
Integumentary system	0.15 (1.21)	0 (0-0)	0.37 (2.18)	0 (0-0)	0.08 (0.62)	0 (0-0)
Other diagnostic or therapeutic procedures	2.48 (5.67)	1 (0-2)	3.17 (4.76)	2 (0-4)	2.26 (5.92)	0 (0-1)
Measurement instruments	0.87 (6.04)	0 (0-0)	0.68 (5.48)	0 (0-0)	0.93 (6.21)	0 (0-0)
Rehabilitation	0.02 (0.15)	0 (0-0)	0.01 (0.12)	0 (0-0)	0.02 (0.16)	0 (0-0)
<i>Age, gender, mortality</i>						
Age	55.9 (26.9)	60 (35-80)	69.3 (20.4)	75 (60-85)	51.6 (27.4)	55 (30-75)
Share of female patients	50%		58%		47%	
Raw inpatient mortality rate	4.65%		14.46%		1.47%	

Annotations: CHOP = Swiss Operation and Procedure Catalogue; ICD = International Classification of Diseases; IMCU = Intermediate Care Units; IUC = Infrastructure Usage Costs. All costs are in Swiss franks and rounded to full numbers. All shares are rounded to two decimals. The number of cases per CHOP chapter are rounded to two decimals and age to one decimal. Labeled sepsis cases follow the explicit coding definition presented in Table 4 in the supplements.

Labeled sepsis cases exhibit more than three times higher total costs than unlabeled cases. Mean costs associated with surgery (operating room cost attributes), the intensive care and intermediate care units, the emergency unit, medical and diagnostic services, and nursing care are all considerably higher for labeled sepsis cases compared to unlabeled cases. Labeled sepsis cases on average receive considerably more surgeries of the respiratory and cardiovascular systems, and of the digestive tract. This is only partially mirrored in Table 2 showing more main diagnoses in ICD chapter J (diseases of the respiratory system) for labeled cases. Patients labeled as sepsis cases are on average roughly 18 years older than unlabeled cases. The raw inpatient mortality rate is almost ten times higher for labeled than for unlabeled cases.

**Table 2: Descriptive results of ICD chapters**

ICD chapter	Total sample (n=264,915)		Labeled sepsis cases (n=64,915)		Unlabeled cases (n=200,000)	
	Number	Share	Number	Share	Number	Share
A / B - Certain infectious and parasitic diseases (A00-B99)	29'653	11%	25'845	40%	3'808	2%
C / D - Neoplasms (C00-D49) and Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism (D50-D89)	22'753	9%	5'407	8%	17'346	9%

E - Endocrine, nutritional and metabolic diseases	4'118	2%	591	1%	3'527	2%
F - Mental, behavioral and neurodevelopmental disorders	16'942	6%	454	1%	16'488	8%
G - Diseases of the nervous system	5'926	2%	621	1%	5'305	3%
H - Diseases of the eye and adnexa (H00-H59) and Diseases of the ear and mastoid process (H60-H95)	2'500	1%	39	0%	2'461	1%
I - Diseases of the circulatory system	24'725	9%	3'569	5%	21'156	11%
J - Diseases of the respiratory system	20'486	8%	7'870	12%	12'616	6%
K - Diseases of the digestive system	21'566	8%	4'837	7%	16'729	8%
L - Diseases of the skin and subcutaneous tissue	3'255	1%	626	1%	2'629	1%
M - Diseases of the musculoskeletal system and connective tissue	23'597	9%	1'297	2%	22'300	11%
N - Diseases of the genitourinary system	19'628	7%	8'670	13%	10'958	5%
O - Pregnancy, childbirth and the puerperium	15'900	6%	292	0%	15'608	8%
P - Certain conditions originating in the perinatal period	6'385	2%	887	1%	5'498	3%
Q - Congenital malformations, deformations and chromosomal abnormalities	1'773	1%	149	0%	1'624	1%
R - Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	6'651	3%	558	1%	6'093	3%
S / T - Injury, poisoning and certain other consequences of external causes (S00-T88)	28'187	11%	3'192	5%	24'995	12%
Z - Factors that influence health status and lead to healthcare utilization	10'757	4%	11	0%	10'746	5%
Not specified	113	0%	0	0%	113	0%

Annotations: ICD = International Classification of Diseases. Labeled sepsis cases follow the explicit coding definition presented in Table 4 in the supplements.

### *Model evaluation*

We present the results of the model evaluation in Table 3. We train our model with positive examples according to the explicit coding strategy (see Table 4 in the supplements for coding strategies). In the second column of Table 3, positive examples in the test set are labeled according to this explicit coding strategy and compared with model predictions for model evaluation. To receive additional insights into the performance of our model, we label positive examples differently in the test set for three more evaluation scenarios. In the third column, we label an example positive according to the implicit coding strategy. These examples might also be positive according to the explicit coding strategy but do not have to be. In the fourth column, positive labels are only given to cases in the test set that are positive according to both the explicit as well as the implicit coding strategy. In the fifth column, examples are labeled positive exclusively according to the implicit coding strategy.

**Table 3: Evaluation metrics**

Evaluation metric	Explicit	Implicit	Explicit and implicit	Only implicit
Recall	<b>0.851</b>	0.842	0.812	0.555
Recall at 10%	0.103	<b>0.117</b>	0.099	0.061
Recall at 20%	0.188	<b>0.192</b>	0.185	0.151
Recall at 30%	0.246	0.246	0.246	<b>0.252</b>
Precision at 10%	0.872	0.790	<b>0.923</b>	0.051
Precision at 20%	0.800	0.648	<b>0.862</b>	0.062
Precision at 30%	0.697	0.555	<b>0.766</b>	0.069
Number of cases in the final sample	64,915	50,372	40,740	9,632
Evaluation scenario	A	B	C	D

Annotations: Coding strategies are presented in Table 4 in the supplements. Highest values per evaluation metric are in bold letters.

Recall is above 0.80 for the three evaluations involving positive examples labeled according to the explicit coding strategy. Recall is highest for evaluation scenario A. For evaluation scenario D, recall still is at 0.555, indicating that the model correctly predicts more than 50% of the examples as TP, although it was trained on a different coding strategy.

Recall at k% is highest for evaluation scenario B for k=10% and k=20%. Recall at k% first is relatively low in evaluation scenario D. Yet it improves strongly and eventually is the highest among the four scenarios for k=30%, albeit at a low level as only 25.2% of TP in the test set are among the top 30% of model predictions.

Scenario D shows the lowest values for precision at k% while Scenario C scores highest. For precision at 10%, for instance, 92.3% of the top 10% of model predictions are true positive examples. This means that the top-k% of predictions are most reliably true positive examples in Scenario C.

## 4 Discussion

We test whether PU learning is an efficient approach to increase the quality of hospital administrative data. Diagnosis codes, for instance, are only actively labeled if they receive a positive class label. Sepsis, commonly underreported in hospital administrative data, serves as application example for identifying positive examples in unlabeled data. We hypothesize that cost data at case level could be useful for identifying positive examples of sepsis in unlabeled data as symptoms of sepsis will always need to be treated, incurring costs, even if sepsis as such was not diagnosed, documented, or coded.

Our results show that the employed PU learning approach (AdaSampling) used with SVM as classifier is effective in correctly identifying positive examples in unseen data. Recall is higher than 0.80 for evaluation scenarios A, B, and C. In these scenarios, all or large parts of the labeled positive examples in the test set follow the same explicit coding strategy as the labeled positive examples in the train set.

Our model even correctly identifies the majority of positive examples of sepsis (55.5%) exclusively following an implicit coding strategy in scenario D.

When looking at the top 10%, 20% and 30% of predictions, our model is less effective at correctly labeling positive examples in these top-k% groups. Evaluation scenarios focusing on the implicit coding strategy perform best in terms of recall at k%. This indicates that there must be considerable number of implicit coding positive examples in the unlabeled data that exhibit very similar cost patterns as the explicit coding positive examples the model was trained on. This is also in line with the relatively high recall for evaluation scenario D discussed earlier.

We find high precision at k% for evaluation scenarios A, B, and C. This means that our model performs well in identifying positive examples with its top predictions in these scenarios.

Schwarzkopf et al. (2023) find low sensitivity when sepsis cases are identified from raw hospital administrative data. Our findings indicate that PU learning can support research using hospital administrative data, e.g., in health economics, health services research, epidemiology, and disease surveillance. Concretely, we suggest that studies could include PU learning in two sensitivity analyses: First, a PU learner could be used to assess data quality and the potential number of positive examples in the unlabeled data. Second, if the first step hints at a high number of positives in the unlabeled data, labels of the top k% should be flipped and the main model should be re-run with the data with new class labels. This is not only relevant for sepsis but also inpatient complications and patient safety indicators estimated from hospital administrative data. Risk-adjustment using co-morbidities identified from hospital administrative data could also be improved with these two PU learning steps. A sepsis-specific implication of the high estimates for precision at k% is that PU learners could help prioritize cases for medical record review: If precision at k% is high, the top-k% predictions should be relabeled with the PU learner and de-prioritized from medical record review.

### *Findings from the literature*

Our results are difficult to compare to other studies as the only healthcare applications we are aware of investigate very different research questions.

### *Limitations*

While the robust PU learning approach AdaSampling has proven reliability for different applications, our findings are not generalizable to all PU learning approaches. Studies using other learning algorithms should be carried out to test what PU learners are best suited for improving hospital administrative data quality. In addition, different classifiers such as random forests, or gradient boosting might also yield different results compared to the SVM classifier we employed in this study.

The implications of our findings are limited by the fact that we can only partially evaluate our model. One possibility to complement model evaluation could be to carry out an empirical Monte Carlo simulation to estimate finite sample performance of PU learners under predefined conditions.

Lastly, PU learning is not limited to sepsis or re-labeling diagnosis codes but could also be a meaningful application for procedure codes, e.g., for robotic-assisted surgery systems (CHOP code 00.99.50 Use of a surgical robot).

## **5 Conclusion**

With PU learning, we propose a novel approach for increasing the quality of hospital administrative data using the identification of sepsis cases in unlabeled data as application example. The implications from our research is that studies using hospital administrative data should employ PU learners to check the quality of their data and, if quality is sub-par, flip class labels for top predictions of the PU learner for a sensitivity analysis. Potentially, this approach could improve the performance of downstream tasks such as risk-adjustment, DRG coding and derivation of DRG cost weights, and estimation of inpatient complication rates or patient safety indicators for (causal) outcomes research.

## Bibliography

- Angus, D.C., Linde-Zwirble, W.T., Lidicker, J., Clermont, G., Carcillo, J., Pinsky, M.R., 2001. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Crit Care Med* 29, 1303–1310. <https://doi.org/10.1097/00003246-200107000-00002>
- Arjannikov, T., Tzanetakis, G., 2021a. An Empirical Investigation of PU Learning for Predicting Length of Stay. *Proceedings - 2021 IEEE 9th International Conference on Healthcare Informatics, ISCHI 2021* 41–47. <https://doi.org/10.1109/ICHI52183.2021.00019>
- Arjannikov, T., Tzanetakis, G., 2021b. Cold-start hospital length of stay prediction using positive-unlabeled learning. *BHI 2021 - 2021 IEEE EMBS International Conference on Biomedical and Health Informatics, Proceedings*. <https://doi.org/10.1109/BHI50953.2021.9508596>
- Bekker, J., Davis, J., 2020. Learning from positive and unlabeled data: a survey. *Mach Learn* 109, 719–760. <https://doi.org/10.1007/S10994-020-05877-5/FIGURES/6>
- Bekker, J., Robberechts, P., Davis, J., 2020. Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data, in: Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis, M., Robardet, C. (Eds.), *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Lecture Notes in Computer Science()*. Springer, Cham, pp. 71–85. <https://doi.org/https://doi.org/10.48550/arXiv.1809.03207>
- Blümel, M., Spranger, A., Achstetter, K., Maresso, A., Busse, R., 2020. Germany: Health system review 2020. *Health Syst Transit* 22, 1–272.
- Buckley, C., Voorhees, E.M., 2000. Evaluating evaluation measure stability. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)* 33–40. <https://doi.org/10.1145/345508.345543>
- Calvo, B., Larrañaga, P., Lozano, J.A., 2007a. Learning Bayesian classifiers from positive and unlabeled examples. *Pattern Recognit Lett* 28, 2375–2384. <https://doi.org/10.1016/J.PATREC.2007.08.003>
- Calvo, B., López-Bigas, N., Furney, S.J., Larrañaga, P., Lozano, J.A., 2007b. A partially supervised classification approach to dominant and recessive human disease gene prediction. *Comput Methods Programs Biomed* 85, 229–237. <https://doi.org/10.1016/J.CMPB.2006.12.003>
- Cerulo, L., Elkan, C., Ceccarelli, M., 2010. Learning gene regulatory networks from only positive and unlabeled data. *BMC Bioinformatics* 11, 1–16. <https://doi.org/10.1186/1471-2105-11-228/FIGURES/13>

- De Pietro, C., Camenzind, P., Sturny, I., Crivelli, L., Edwards-Garavoglia, S., Spranger, A., Wittenbecher, F., Quentin, W., 2015. Switzerland: Health system review. *Health Syst Transit* 17, 1–288.
- Denis, F., Gilleron, R., Letouzey, F., 2005. Learning from positive and unlabeled examples. *Theor Comput Sci* 348, 70–83. <https://doi.org/10.1016/J.TCS.2005.09.007>
- Elkan, C., Noto, K., 2008. Learning classifiers from only positive and unlabeled data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 213–220. <https://doi.org/10.1145/1401890.1401920>
- Fleischmann-Struzek, C., Mikolajetz, A., Schwarzkopf, D., Cohen, J., Hartog, C.S., Pletz, M., Gastmeier, P., Reinhart, K., 2018. Challenges in assessing the burden of sepsis and understanding the inequalities of sepsis outcomes between National Health Systems: secular trends in sepsis and infection incidence and mortality in Germany. *Intensive Care Med* 44, 1826–1835. <https://doi.org/10.1007/S00134-018-5377-4/TABLES/2>
- Fleischmann-Struzek, C., Rudd, K., 2023. Challenges of assessing the burden of sepsis. *Med Klin Intensivmed Notfmed* 118, 68–74. <https://doi.org/10.1007/S00063-023-01088-7/FIGURES/2>
- Freitas, A., Gaspar, J., Rocha, N., Marreiros, G., Da Costa-Pereira, A., 2014. Quality in Hospital Administrative Databases. *Applied Mathematics & Information Sciences* 8, 1–6. <https://doi.org/10.12785/AMIS/081L01>
- Jaskie, K., Spanias, A., 2019. Positive and Unlabeled Learning Algorithms and Applications: A Survey. *10th International Conference on Information, Intelligence, Systems and Applications, IISA 2019*. <https://doi.org/10.1109/IISA.2019.8900698>
- Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. *Artif Intell* 97, 273–324. [https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/10.1016/S0004-3702(97)00043-X)
- Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P.S., 2003. Building text classifiers using positive and unlabeled examples. *Proceedings - IEEE International Conference on Data Mining, ICDM* 179–186. <https://doi.org/10.1109/ICDM.2003.1250918>
- Maass, C., Kuske, S., Lessing, C., Schrappe, M., 2015. Are administrative data valid when measuring patient safety in hospitals? A comparison of data collection methods using a chart review and administrative data. *International Journal for Quality in Health Care* 27, 305–313. <https://doi.org/10.1093/INTQHC/MZV045>
- Manning, C., Raghavan, P., Schütze, H., 2009. Chapter 8: Evaluation in information retrieval, in: *Introduction to Information Retrieval*. Cambridge University Press, pp. 151–175.
- Mellhammar, L., Wollter, E., Dahlberg, J., Donovan, B., Olsén, C.J., Wiking, P.O., Rose, N., Schwarzkopf, D., Friedrich, M., Fleischmann-Struzek, C., Reinhart, K., Linder, A., 2023. Estimating Sepsis Incidence Using Administrative Data and Clinical Medical Record Review.

<https://doi.org/10.1001/JAMANETWORKOPEN.2023.31168>

- Rhee, C., Gohil, S., Klompas, M., 2014. Regulatory Mandates for Sepsis Care — Reasons for Caution. *New England Journal of Medicine* 370, 1673–1676. [https://doi.org/10.1056/NEJMP1400276/SUPPL\\_FILE/NEJMP1400276\\_DISCLOSURES.PDF](https://doi.org/10.1056/NEJMP1400276/SUPPL_FILE/NEJMP1400276_DISCLOSURES.PDF)
- Rhee, C., Murphy, M. V., Li, L., Platt, R., Klompas, M., 2015. Comparison of Trends in Sepsis Incidence and Coding Using Administrative Claims Versus Objective Clinical Data. *Clinical Infectious Diseases* 60, 88–95. <https://doi.org/10.1093/CID/CIU750>
- Schwarzkopf, D., Rose, N., Fleischmann-Struzek, C., Boden, B., Dorow, H., Edel, A., Friedrich, M., Gonnert, F.A., Götz, J., Gründling, M., Heim, M., Holbeck, K., Jaschinski, U., Koch, C., Künzer, C., Le Ngoc, K., Lindau, S., Mehlmann, N.B., Meschede, J., Meybohm, P., Quart, D., Putensen, C., Sander, M., Schewe, J.C., Schlattmann, P., Schmidt, G., Schneider, G., Spies, C., Steinsberger, F., Zacharowski, K., Zinn, S., Reinhart, K., 2023. Understanding the biases to sepsis surveillance and quality assurance caused by inaccurate coding in administrative health data. *Infection* 1, 1–15. <https://doi.org/10.1007/S15010-023-02091-Y/FIGURES/3>
- Singer, M., Deutschman, C.S., Seymour, C., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G.R., Chiche, J.D., Coopersmith, C.M., Hotchkiss, R.S., Levy, M.M., Marshall, J.C., Martin, G.S., Opal, S.M., Rubenfeld, G.D., Poll, T. Der, Vincent, J.L., Angus, D.C., 2016. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 315, 801–810. <https://doi.org/10.1001/JAMA.2016.0287>
- Tiru, B., DiNino, E.K., Orenstein, A., Mailloux, P.T., Pesaturo, A., Gupta, A., McGee, W.T., 2015. The Economic and Humanistic Burden of Severe Sepsis. *Pharmacoeconomics* 33, 925–937. <https://doi.org/10.1007/S40273-015-0282-Y>
- Wang, C., Ding, C., Meraz, R.F., Holbrook, S.R., 2006. PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. *Bioinformatics* 22, 2590–2596. <https://doi.org/10.1093/BIOINFORMATICS/BTL441>
- Yang, P., Liu, W., Yang, J., 2017. Positive unlabeled learning via wrapper-based adaptive sampling. *IJCAI International Joint Conference on Artificial Intelligence* 0, 3273–3279. <https://doi.org/10.24963/IJCAI.2017/457>
- Yang, P., Ormerod, J.T., Liu, W., Ma, C., Zomaya, A.Y., Yang, J.Y.H., 2019. AdaSampling for positive-unlabeled and label noise learning with bioinformatics applications. *IEEE Trans Cybern* 49, 1932–1943. <https://doi.org/10.1109/TCYB.2018.2816984>
- Yu, H., Zhai, C., Han, J., 2003. Text classification from positive and unlabeled documents. *Proceedings of the twelfth international conference on Information and knowledge management* 232–239. <https://doi.org/10.1145/956863.956909>



- Zhang, J., Wang, Z., Meng, J., Tan, Y.P., Yuan, J., 2019. Boosting positive and unlabeled learning for anomaly detection with multi-features. *IEEE Trans Multimedia* 21, 1332–1344. <https://doi.org/10.1109/TMM.2018.2871421>
- Zhou, J., Lu, Xiaoyu, Chang, W., Wan, C., Lu, Xiongbin, Zhang, C., Cao, S., 2022. PLUS: Predicting cancer metastasis potential based on positive and unlabeled learning. *PLoS Comput Biol* 18, e1009956. <https://doi.org/10.1371/JOURNAL.PCBI.1009956>

## Supplementary material

**Table 4: Explicit and implicit sepsis codes**

	<b>Explicit Sepsis-1 coding</b>	<b>Implicit coding (Angus definition)</b>
Explicit sepsis-1 ICD-10-codes	A021; A200; A207; A217; A227; A241; A267; A282; A327; A391; A392; A393; A394; A40; A41; A427; A483; A499; A548; B007; B376; B377; B49; P36; R572; R650; R651	-
ICD-10-codes for infection	-	A00; A01; A02; A03; A04; A05; A06; A07; A08; A09; A15; A16; A17; A18; A19; A20; A21; A22; A23; A24; A25; A26; A27; A28; A32; A36; A37; A38; A39; A40; A41; A42; A43; A44; A46; A48; A49; A50; A54; A55; A56; A59; A65; A690; A691; A692; A698; A699; A74; A75; A77; A78; A79; A80; A81; A83; A84; A85; A86; A87; A88; A89; A90; A91; A92; A93; A94; A95; A96; A97; A98; A99; B00; B01; B02; B03; B04; B05; B06; B07; B08; B09; B25; B26; B27; B33; B34; B37; B38; B39; B40; B41; B42; B43; B44; B45; B46; B47; B48; B49; B50; B51; B52; B53; B54; B55; B58; B60; B64; B67; B95; B96; B97; B98; B99; G00; G01; G02; G03; G04; G05; G06; G07; G08; H050; H602; H700; I32; I33; I38; I39; I40; I41; I80; I981; J01; J02; J03; J04; J05; J06; J09; J10; J11; J12; J13; J14; J15; J16; J17; J18; J20; J21; J22; J36; J390; J391; J440; J441; J85; J86; K35; K36; K37; K5702; K5703; K5712; K5713; K5722; K5723; K5732; K5733; K5742; K5743; K5752; K5753; K5782; K5783; K5792; K5793; K61; K630; K631; K65; K67; K750; K751; K770; K810; L02; L03; L04; L05; L08; M00; M01; M86; N10; N151; N159; N30; N34; N390; N41; N45; N482; N49; N61; N70; N71; N72; N73; N74; N75; N76; N77; N980; O030; O035; O040; O045; O050; O055; O060; O065; O070; O075; O080; O23; O411; O753; O85; O86; O883; O91; O98; P23; P240; P248; P249; P35; P36; P37; P38; P39; P77; P781; R572; R650; R651; T802; T814; T826; T827; T835; T836; T845; T846; T847; T857; T880; U6900; U6940
ICD-10-codes for organ dysfunction	-	D65; D688; D689; D695; D696; E872; F05; G931; G934; I959; J80; J960; J969; J984; K720; K727; K762; K763; N17; N19; R060; R068; R40; R572; R578; R579; R651

Notes: If a code was present for a case, it was counted as a positive example. If the code was not present, the case was counted as a negative or rather unlabeled example. For the Angus definition, both at least one code from the ICD-10-codes for infection and at least one code from the ICD-10-codes for organ dysfunction needed to be present for a case to be counted as a positive example. We considered both primary and secondary diagnoses. If an ICD-10-code is listed with only two or three characters, this means that all subordinate codes (i.e., all more detailed ICD-10-codes with three or four characters) were considered (if available for the particular ICD-10-code). Definitions are based on Schwarzkopf et al. (2023) and Angus et al. (2001).

**Table 5: Included attributes PU data classifier could learn from**

Category	Attribute
Hospital costs at case level	Total costs
	Pharmaceuticals
	Blood and blood products
	Medical material
	Implants
	Medical, diagnostic and therapeutic third-party services (excl. doctors' fees)
	Doctor's fees (not subject to social security contributions)
	Medical fees, hospital doctors (subject to social security contributions)
	Medical fees, attending physicians (subject to social security contributions)
	Patient transportation by third parties
	Other patient-related third-party services
	Other expenses for patients
	Patient administration, overheads excl. IUC
	Patient administration, IUC
	Operating room, overheads excl. IUC
	Operating room, IUC
	Operating room doctors - activities 6a, overheads excl. IUC
	Operating room doctors - activities 6a, IUC
	Anesthesia, overheads excl. IUC
	Anesthesia, IUC
	Intensive care unit, overheads excl. IUC
	Intensive care unit, IUC
	Intensive care unit physicians - activities 6b1, overheads excl. IUC
	Intensive care unit physicians - activities 6b1, IUC
	Intermediate Care Units (IMCU), overheads excl. IUC
	Intermediate Care Units (IMCU), IUC
	IMCU physicians - activities 6b2, overheads excl. IUC
	IMCU physicians - activities 6b2, IUC
	Emergency, overheads excl. IUC
	Emergency, IUC
	Emergency physician services - activities 6b3, overheads excl. IUC
	Emergency physician services - activities 6b3, IUC
	Imaging procedures, overheads excl. IUC
	Imaging procedures, IUC
	Delivery room, overheads excl. IUC
	Delivery room, IUC
	Delivery room doctors - activities 6b4, overheads excl. IUC
	Delivery room doctors - activities 6b4, IUC
	Nuclear medicine and radiation oncology, overheads excl. IUC
	Nuclear medicine and radiation oncology, IUC
	Laboratory, overheads excl. IUC
	Laboratory, IUC
Dialysis, overheads excl. IUC	
Dialysis, IUC	
Physicians, activities 1-5, overheads excl. IUC	
Physicians, activities 1-5, IUC	
Physiotherapy, overheads excl. IUC	
Physiotherapy, IUC	
Occupational therapy, overheads excl. IUC	
Occupational therapy, IUC	
Speech therapy, overheads excl. IUC	
Speech therapy, IUC	

Category	Attribute
	Non-medical therapies and consultations, overheads excl. IUC
	Non-medical therapies and consultations, IUC
	Medical and therapeutic diagnostics, overheads excl. IUC
	Medical and therapeutic diagnostics, IUC
	Medical and therapeutic diagnostics doctors - activities 6b5, overheads excl.
	Medical and therapeutic diagnostics - activities 6b5, IUC
	Nursing, overheads excl. IUC
	Nursing care, IUC
	Hotel rooms, overheads excl. IUC
	Hotel rooms, IUC
	Hotel kitchen, overheads excl. IUC
	Hotel kitchen, IUC
	Hotel service, overheads excl. IUC
	Hotel service, IUC
	Other service providers, overheads excl. IUC
	Other service providers, IUC
	Pathology, overheads excl. IUC
	Pathology, IUC
	Rescue and ambulance service (secondary transport only), overheads excl. IUC
	Rescue and ambulance service (secondary transports only), IUC
Aggregated medical information	ICD chapter of main diagnosis
	Number of procedures per CHOP chapter

Annotations: IUC = Infrastructure Usage Costs; ICD = International Classification of Diseases; CHOP = Swiss Operation and Procedure Catalogue. All attributes of the category hospital costs at case level are in Swiss francs and continuous variables with a theoretical range between 0 and positive infinity. ICD chapters are dummy variables. The number of procedures per CHOP chapter are natural numbers including 0. Overall, there are 23 ICD (21 relevant for main diagnoses) and 19 CHOP chapters.

**Table 6: Descriptive results of cost attributes, procedure chapters, age, gender, and inpatient mortality**

Attribute	Total sample (n=264,915)		Labeled sepsis cases (n=64,915)		Unlabeled cases (n=200,000)	
	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )
Total costs	17895 (40575)	8515 (4837- 16610)	36342 (71718)	14731 (7720- 34426)	11907 (19104)	7418 (4240- 13363)
<i>Variable direct costs [CHF]</i>						
Pharmaceuticals	518 (3183)	89 (19-252)	1398 (5783)	261 (94- 817)	232 (1494)	63 (10-168)
Blood and blood products	227 (3055)	0 (0-0)	763 (5935)	0 (0-0)	53 (896)	0 (0-0)
Medical material	587 (2066)	77 (1-418)	1009 (3448)	108 (17- 593)	450 (1311)	66 (0-381)
Implants	436 (2726)	0 (0-0)	385 (3613)	0 (0-0)	453 (2367)	0 (0-0)
Medical, diagnostic and therapeutic third-party services (excl. doctors' fees)	258 (1182)	0 (0-127)	597 (1942)	1 (0-476)	148 (760)	0 (0-66)
Doctor's fees (not subject to social security contributions)	265 (1342)	0 (0-0)	272 (2186)	0 (0-0)	262 (913)	0 (0-0)
Medical fees, hospital doctors (subject to social security contributions)	22 (197)	0 (0-0)	31 (258)	0 (0-0)	19 (172)	0 (0-0)
Medical fees, attending physicians (subject to social security contributions)	17 (182)	0 (0-0)	21 (206)	0 (0-0)	16 (174)	0 (0-0)
Patient transportation by third parties	68 (401)	0 (0-0)	127 (608)	0 (0-0)	48 (303)	0 (0-0)
Other patient-related third-party services	15 (493)	0 (0-0)	35 (667)	0 (0-0)	9 (421)	0 (0-0)
Other expenses for patients	11 (203)	0 (0-0)	26 (333)	0 (0-0)	6 (137)	0 (0-0)
<i>Fixed indirect costs [CHF]</i>						
Patient administration, overheads excl. IUC	196 (199)	172 (112- 239)	211 (204)	184 (119- 244)	191 (197)	167 (111- 237)
Patient administration, IUC	17 (24)	11 (6-21)	18 (29)	11 (7-20)	17 (22)	11 (6-21)
Operating room, overheads excl. IUC	711 (2174)	0 (0-830)	982 (3765)	0 (0-0)	623 (1275)	0 (0-934)
Operating room, IUC	181 (516)	0 (0-197)	233 (833)	0 (0-1)	164 (355)	0 (0-225)
Operating room doctors - activities 6a, overheads excl. IUC	238 (903)	0 (0-0)	342 (1387)	0 (0-0)	204 (672)	0 (0-34)
Operating room doctors - activities 6a, IUC	13 (74)	0 (0-0)	18 (106)	0 (0-0)	11 (59)	0 (0-0)
Anesthesia, overheads excl. IUC	602 (1650)	0 (0-758)	850 (2759)	0 (0-493)	521 (1054)	0 (0-802)
Anesthesia, IUC	58 (160)	0 (0-69)	79 (263)	0 (0-39)	50 (106)	0 (0-74)
Intensive care unit, overheads excl. IUC	2165 (14418)	0 (0-0)	7592 (27602)	0 (0-3274)	403 (3925)	0 (0-0)
Intensive care unit, IUC	204 (1346)	0 (0-0)	714 (2573)	0 (0-271)	38 (374)	0 (0-0)
Intensive care unit physicians - activities 6b1, overheads excl. IUC	411 (3236)	0 (0-0)	1418 (6225)	0 (0-139)	84 (928)	0 (0-0)
Intensive care unit physicians - activities 6b1, IUC	17 (167)	0 (0-0)	56 (310)	0 (0-0)	4 (70)	0 (0-0)
IMCU, overheads excl. IUC	290 (3050)	0 (0-0)	799 (5671)	0 (0-0)	125 (1331)	0 (0-0)
Intermediate Care Units (IMCU), IUC	21 (246)	0 (0-0)	55 (455)	0 (0-0)	9 (111)	0 (0-0)

Attribute	Total sample (n=264,915)		Labeled sepsis cases (n=64,915)		Unlabeled cases (n=200,000)	
	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )
IMCU physicians - activities 6b2, overheads excl. IUC	41 (445)	0 (0-0)	107 (803)	0 (0-0)	19 (225)	0 (0-0)
IMCU physicians - activities 6b2, IUC	2 (26)	0 (0-0)	5 (46)	0 (0-0)	1 (15)	0 (0-0)
Emergency, overheads excl. IUC	196 (330)	0 (0-323)	352 (369)	307 (87-479)	145 (300)	0 (0-235)
Emergency, IUC	20 (36)	0 (0-31)	36 (41)	26 (6-53)	15 (33)	0 (0-22)
Emergency physician services - activities 6b3, overheads excl. IUC	94 (188)	0 (0-133)	167 (227)	68 (0-280)	70 (166)	0 (0-44)
Emergency physician services – activities 6b3, IUC	4 (12)	0 (0-1)	7 (15)	0 (0-7)	3 (10)	0 (0-0)
Imaging procedures, overheads excl. IUC	293 (731)	55 (0-312)	657 (1179)	297 (86-744)	175 (446)	0 (0-176)
Imaging procedures, IUC	81 (212)	13 (0-82)	179 (328)	76 (23-199)	50 (142)	0 (0-46)
Delivery room, overheads excl. IUC	105 (604)	0 (0-0)	7 (195)	0 (0-0)	137 (683)	0 (0-0)
Delivery room, IUC	14 (89)	0 (0-0)	1 (28)	0 (0-0)	18 (101)	0 (0-0)
Delivery room doctors - activities 6b4, overheads excl. IUC	18 (149)	0 (0-0)	2 (62)	0 (0-0)	23 (168)	0 (0-0)
Delivery room doctors - activities 6b4, IUC	1 (12)	0 (0-0)	0 (7)	0 (0-0)	1 (13)	0 (0-0)
Nuclear medicine and radiation oncology, overheads excl. IUC	22 (303)	0 (0-0)	34 (371)	0 (0-0)	18 (277)	0 (0-0)
Nuclear med. and radiation oncology, IUC	8 (112)	0 (0-0)	12 (136)	0 (0-0)	7 (102)	0 (0-0)
Laboratory, overheads excl. IUC	520 (1878)	108 (0-405)	1477 (3459)	570 (251-1270)	209 (627)	58 (0-210)
Laboratory, IUC	61 (232)	11 (0-44)	175 (428)	61 (23-147)	25 (82)	6 (0-23)
Dialysis, overheads excl. IUC	70 (1107)	0 (0-0)	239 (2137)	0 (0-0)	15 (358)	0 (0-0)
Dialysis, IUC	8 (125)	0 (0-0)	25 (238)	0 (0-0)	2 (45)	0 (0-0)
Physicians, activities 1-5, overh. excl. IUC	1411 (3168)	790 (216-1613)	2561 (5343)	1443 (738-2733)	1038 (1860)	623 (112-1289)
Physicians, activities 1-5, IUC	97 (234)	38 (2-104)	170 (360)	76 (23-175)	73 (169)	28 (0-85)
Physiotherapy, overheads excl. IUC	267 (1090)	0 (0-208)	609 (1826)	185 (0-598)	156 (665)	0 (0-114)
Physiotherapy, IUC	35 (135)	0 (0-25)	73 (190)	19 (0-74)	22 (109)	0 (0-13)
Occupational therapy, overheads excl. IUC	43 (431)	0 (0-0)	62 (554)	0 (0-0)	36 (383)	0 (0-0)
Occupational therapy, IUC	5 (62)	0 (0-0)	7 (70)	0 (0-0)	5 (59)	0 (0-0)
Speech therapy, overheads excl. IUC	18 (247)	0 (0-0)	43 (398)	0 (0-0)	10 (171)	0 (0-0)
Speech therapy, IUC	2 (24)	0 (0-0)	4 (41)	0 (0-0)	1 (14)	0 (0-0)
Non-medical therapies and consultations, overheads excl. IUC	148 (911)	0 (0-31)	257 (812)	3 (0-225)	112 (938)	0 (0-13)
Non-med. therapies and consultations, IUC	15 (116)	0 (0-3)	26 (97)	1 (0-20)	11 (121)	0 (0-1)
Medical and therapeutic diagnostics, overheads excl. IUC	134 (613)	0 (0-0)	234 (823)	0 (0-121)	101 (523)	0 (0-0)
Medical and therapeutic diagnostics, IUC	37 (155)	0 (0-0)	64 (212)	0 (0-31)	28 (129)	0 (0-0)

Attribute	Total sample (n=264,915)		Labeled sepsis cases (n=64,915)		Unlabeled cases (n=200,000)	
	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )
Medical and therapeutic diagnostics doctors - activities 6b5, overheads excl.	71 (412)	0 (0-0)	155 (684)	0 (0-13)	44 (265)	0 (0-0)
Medical and therapeutic diagnostics - activities 6b5, IUC	3 (22)	0 (0-0)	7 (33)	0 (0-0)	2 (16)	0 (0-0)
Nursing care, overheads excl. IUC	3705 (8127)	1582 (720-3719)	7158 (12458)	3670 (1616-8028)	2584 (5656)	1282 (630-2644)
Nursing care, IUC	256 (625)	104 (36-252)	462 (945)	210 (79-500)	189 (458)	85 (31-193)
Hotel rooms, overheads excl. IUC	371 (809)	183 (81-388)	670 (1235)	355 (178-723)	274 (578)	150 (67-299)
Hotel rooms, IUC	267 (792)	118 (32-280)	473 (1027)	225 (72-517)	201 (686)	97 (23-220)
Hotel kitchen, overheads excl. IUC	452 (767)	247 (107-519)	776 (993)	496 (258-947)	347 (644)	199 (88-394)
Hotel kitchen, IUC	63 (239)	30 (9-70)	104 (157)	58 (21-125)	50 (259)	24 (7-55)
Hotel service, overheads excl. IUC	95 (375)	0 (0-71)	152 (447)	0 (0-120)	76 (347)	0 (0-58)
Hotel service, IUC	7 (31)	0 (0-3)	9 (32)	0 (0-6)	6 (31)	0 (0-3)
Other serv. providers, overheads excl. IUC	41 (224)	0 (0-22)	73 (277)	3 (0-51)	30 (203)	0 (0-16)
Other service providers, IUC	13 (97)	0 (0-3)	25 (155)	0 (0-8)	9 (67)	0 (0-2)
Pathology, overheads excl. IUC	43 (294)	0 (0-0)	77 (437)	0 (0-0)	32 (229)	0 (0-0)
Pathology, IUC	8 (60)	0 (0-0)	15 (95)	0 (0-0)	6 (42)	0 (0-0)
Rescue/ ambulance service (secondary transport), overheads excl. IUC	13 (139)	0 (0-0)	26 (190)	0 (0-0)	9 (118)	0 (0-0)
Rescue/ ambulance service (secondary transport), IUC	1 (18)	0 (0-0)	3 (24)	0 (0-0)	1 (16)	0 (0-0)
<i>CHOP chapters: Surgeries of the...</i>						
Procedures not classified elsewhere	0.25 (0.78)	0 (0-0)	0.31 (0.94)	0 (0-0)	0.23 (0.71)	0 (0-0)
Nervous system	0.08 (0.46)	0 (0-0)	0.08 (0.58)	0 (0-0)	0.09 (0.41)	0 (0-0)
Endocrine system	0 (0.07)	0 (0-0)	0 (0.05)	0 (0-0)	0 (0.08)	0 (0-0)
Eye	0.02 (0.25)	0 (0-0)	0 (0.07)	0 (0-0)	0.02 (0.29)	0 (0-0)
Ears	0.01 (0.1)	0 (0-0)	0 (0.06)	0 (0-0)	0.01 (0.11)	0 (0-0)
Nose, mouth, throat	0.04 (0.31)	0 (0-0)	0.02 (0.24)	0 (0-0)	0.04 (0.33)	0 (0-0)
Respiratory system	0.08 (0.54)	0 (0-0)	0.22 (0.95)	0 (0-0)	0.03 (0.29)	0 (0-0)
Cardiovascular system	0.3 (1.65)	0 (0-0)	0.67 (2.82)	0 (0-0)	0.18 (0.98)	0 (0-0)
Hematopoietic and lymphatic system	0.02 (0.19)	0 (0-0)	0.04 (0.25)	0 (0-0)	0.02 (0.17)	0 (0-0)
Digestive tract	0.29 (1.25)	0 (0-0)	0.67 (2.22)	0 (0-0)	0.17 (0.65)	0 (0-0)
Urinary organs	0.11 (0.47)	0 (0-0)	0.22 (0.67)	0 (0-0)	0.07 (0.37)	0 (0-0)
Male sexual organs	0.01 (0.14)	0 (0-0)	0.01 (0.16)	0 (0-0)	0.02 (0.14)	0 (0-0)
Female sexual organs	0.04 (0.31)	0 (0-0)	0.01 (0.16)	0 (0-0)	0.05 (0.35)	0 (0-0)
Obstetric procedures	0.1 (0.48)	0 (0-0)	0 (0.08)	0 (0-0)	0.13 (0.54)	0 (0-0)
Musculoskeletal system	0.28 (1.06)	0 (0-0)	0.21 (1.23)	0 (0-0)	0.3 (1)	0 (0-0)
Integumentary system	0.15 (1.21)	0 (0-0)	0.37 (2.18)	0 (0-0)	0.08 (0.62)	0 (0-0)

Attribute	Total sample (n=264,915)		Labeled sepsis cases (n=64,915)		Unlabeled cases (n=200,000)	
	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )	Mean (SD)	Median (25 <sup>th</sup> – 75 <sup>th</sup> )
Other diagnostic or therapeutic procedures	2.48 (5.67)	1 (0-2)	3.17 (4.76)	2 (0-4)	2.26 (5.92)	0 (0-1)
Measurement instruments	0.87 (6.04)	0 (0-0)	0.68 (5.48)	0 (0-0)	0.93 (6.21)	0 (0-0)
Rehabilitation	0.02 (0.15)	0 (0-0)	0.01 (0.12)	0 (0-0)	0.02 (0.16)	0 (0-0)
<i>Age, gender, mortality</i>						
Age	55.91 (26.94)	60 (35-80)	69.28 (20.38)	75 (60-85)	51.57 (27.38)	55 (30-75)
Share of female patients	50%		58%		47%	
Inpatient mortality rate	4.65%		14.46%		1.47%	

Annotation: CHOP = Swiss Operation and Procedure Catalogue; ICD = International Classification of Diseases; IMCU = Intermediate Care Units; IUC = Infrastructure Usage Costs. All costs are in Swiss francs and rounded to full numbers. All shares are rounded to two decimals. CHOP chapters and age are rounded to two decimals. Labeled sepsis cases follow the explicit coding definition presented in Table 4 in the supplements.