Clinical Epidemiology

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Data-Driven Algorithms for Classification of In- and Outpatients in the Danish National Patient Register

Ann-Sophie Buchardt[1], Pi Vejsig Madsen[1], Andreas Jensen[1,2]

[1]Mary Elizabeth's Hospital, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark; [2]Department of Paediatrics and Adolescent Medicine, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark

Correspondence: Ann-Sophie Buchardt, Rigshospitalet, Blegdamsvej 9, København Ø, 2100, Denmark, Email ann-sophie.buchardt@regionh.dk

Purpose: The Danish National Patient Register (DNPR) is an important data source for research providing detailed information on all hospital contacts in Denmark. With the transition from the second version of the DNPR (DNPR2) to the third version (DNPR3) in early 2019, the patient type variable (inpatient, elective outpatient, acute outpatient) was removed. This study proposes and evaluates algorithms to classify hospital contacts into these categories in DNPR3, aiming for consensus in data interpretation for researchers using Danish registries.

Patients and Methods: We analyzed somatic public hospital contacts in Denmark from 2017 to 2020, with 20,882,018 unique contacts in DNPR2 and 27,694,584 in DNPR3. Several classification algorithms were developed and assessed, including department-based, contact-based, and hybrid methods, to infer patient types in DNPR3 based on contact features, such as duration and admission type. In DNPR3, where the true patient type is unknown, proxy labels were used to train classification algorithms.

Results: Compared to the true patient type variable in DNPR2, our department-based classifier showed high positive predictive values (PPVs) and sensitivities in DNPR2 with PPVs ranging from 95.6 to 99.5 and sensitivities ranging from 94.1 to 99.6 across patient types. The hybrid approach showed improved PPVs and sensitivities for acute (PPV = 97.3, sensitivity = 96.8) and elective (PPV = 99.8, sensitivity = 99.9) outpatients. In both DNPR2 and DNPR3 high agreement between contact-based classification algorithms was obtained indicating robustness in our classification methods which suggests the presence of inherent patterns in the data.

Conclusion: Our study shows that all presented classification methods are suitable for categorizing patient types in DNPR2 depending on the available data and furthermore demonstrated robustness, supporting their suitability for classification in DNPR3. Future research should explore advanced techniques and comprehensive department classification for enhanced accuracy and applicability.

Keywords: classification algorithms, register data, patient types, hospital contacts, hospital departments, unsupervised learning

Introduction

In Denmark, health data are registered for general practitioner and other medical specialist visits, dispensed prescribed medication at pharmacies, and hospital contacts. The data are stored in national health registers, which contain information on diseases, treatments, and other aspects of the healthcare system.[1,2]

The Danish National Patient Register (DNPR) has recorded information about hospital contacts since 1977. The DNPR is a valuable tool for epidemiological research, renowned for its content, data quality, and research potential.[3]

In 2019 a major update of DNPR was carried out, implementing trajectory-oriented registration, which bundles contacts, diagnoses, and procedures related to a patient's clinical treatment.[4] This is an improvement compared to the previous version, DNPR2, where outpatient trajectories were organized as one main contact with a varying number of attached visits. The only visit-specific information available was the date. In DNPR3, each outpatient visit is now defined by its own record, with detailed information on diagnoses, contact duration, and other variables.

Received: 14 October 2024
Accepted: 17 January 2025
Published: 21 February 2025

However, a notable shortcoming of DNPR3 is the fact that the patient type variable no longer exists. The patient type variable classified hospital contacts as inpatient, elective outpatient, or acute outpatient. The information is registered by healthcare professionals, but the information is not available in DNPR3. This omission has resulted in challenges for research using Danish hospital data such as epidemiological studies that assess the risk of specific types of hospital contacts, eg related to certain diagnoses; in DNPR3 it is no longer obvious how to discern inpatient contacts from elective outpatient contacts, which are generally less severe.[5,6]

This study proposes and evaluates algorithms to classify hospital contacts in DNPR3 as inpatient, elective outpatient, or acute outpatient, aiming for a consensus in data interpretation among researchers within the field of register-based research in Denmark.

The Danish Health Data Authority (Danish: Sundhedsdatastyrelsen; SDS) initially suggested defining inpatient contacts as those lasting at least 12 hours. This criterion is used, for example, by Statistics Denmark in publicly available aggregate data on in- and outpatient contacts.[7] The threshold was later revised to 6 hours, accompanied by further criteria based on other available variables. The most recent SDS recommendation defined inpatients as contacts lasting at least 6 hours, elective outpatients as planned contacts lasting less than one hour, and acute outpatients as acute contacts lasting less than 6 hours with additional criteria on the cause of the contacts.[8] However, this proposal did not label all contacts. For instance, a planned contact lasting 2 hours would not fall into any of the categories. Indeed, from the introduction of DNPR3 in early 2019 until the end of 2020, 36% of all hospital contacts fell outside these categories. This presents limitations for studies broadly focused on health care utilization. An alternative method, also suggested by SDS, is to categorize contacts according to the department (ie, emergency department or inpatient ward).[8] However, this approach requires vast manual effort and subject matter knowledge to generalize at the national level.

These suggested definitions are no longer maintained, and no specific criteria are recommended by SDS.

The overarching challenge is that while DNPR2 includes a gold standard for classifying contacts into patient types, it does not include satisfactory data on contact duration. On the other hand, DNPR3 includes the required data on contact duration but not a gold standard for patient type classification – indeed, this is the motivation for the present study.

This paper proposes several data-driven definitions for classifying patient types based on contact-level information and revisits a previously suggested algorithm that classifies contacts based on department-level information.[9] Throughout the paper, we distinguish between three types of classification methods. Contact-based classification methods rely on specific characteristics of the contact, such as the duration, to determine the patient type. Department-based classification methods classify all contacts within a department according to the designated department type. For example, if a department is classified as serving inpatients, all contacts within that department are labeled as inpatient contacts. Finally, a hybrid classification method combines contact- and department-level information. A common feature of previous proposals is their straightforward interpretation, as they are stated in terms of variables available in DNPR3.[8,9] These classification algorithms can be expressed as decision trees, which are known for their interpretability.[10] In line with this preference for simplicity, all classification algorithms proposed in this paper are also presented as decision trees.

The paper is organized as follows: In the Methods section, the data structures of DNPR2 and DNPR3 are first introduced, followed by a brief introduction to the statistical methods. The classification approaches are then presented in three main groups: first three department-based approaches using DNPR2 data, then a hybrid approach also incorporating individual contact-level data, also based on DNPR2, and, finally, three contact-level approaches using DNPR2 and DNPR3 data.[9] We present the results of these classification approaches, and the paper concludes with a discussion of their applicability.

Materials and Methods

Data Source

This study utilized data from the DNPR2 and DNPR3 registers, which cover all somatic public hospital contacts in Denmark. Both DNPR2 and DNPR3 encompass multiple tables, each containing a collection of variables and a common identifier. In the supplementary material (section A.2) we provide a detailed description of the data using the SDS

terminology.[11] We considered the raw contact data from DNPR2 and DNPR3, meaning that adjacent or overlapping contacts were not joined.

All contacts at public somatic hospitals were included, as such contacts at private hospitals, hospices, and psychiatric hospitals were not included, see Table A.1 in the supplementary material for a complete list of included hospitals.

DNPR2

In DNPR2, contacts starting between 1 January 2017 and 31 December 2018 were included. The start date for inpatients was the admission date, for elective outpatients, it was the record start date, and for acute outpatients, it was the arrival date. To align with the DNPR3 structure and to improve comparability, each visit within an elective outpatient record was considered a separate hospital contact.

The full DNPR2 data set used for the department-based approaches contained 1774 unique department codes and 20,882,018 unique contacts. The distribution of true patient type was 2,405,603 (12%) acute outpatients, 15,739,859 (75%) elective outpatients, and 2,736,556 (13%) inpatients. The DNPR2 data set was split into 70% training and 30% validation sets, with 1761 and 1726 unique department codes, and 14,617,479 and 6,264,539 contacts, respectively. We refer to these data sets as the *full DNPR2 training data* and *full DNPR2 validation data*, respectively.

When applying the contact-based approaches further exclusion criteria were necessary. In order to best approximate the duration of elective outpatient visits, we excluded elective outpatient records with more than one registered contact, as well as elective outpatient records lasting overnight, since this is conceptually unreasonable for the contact type. Furthermore, elective outpatient contacts registered with start timestamp "00:00" or end timestamp "00:00" or '23:00' were excluded since these timestamps were deemed unreliable, see for example Figure A.1. Finally, we excluded contacts with a negative duration.

The reduced DNPR2 data set used for the contact-based approaches contained 1294 unique department codes and 6,826,082 unique contacts. The distribution of true patient type was 2,404,132 (35%) acute outpatients, 1,685,401 (25%) elective outpatients, and 2,736,549 (40%) inpatients. The reduced DNPR2 data set was split into 70% training and 30% validation sets, with 1277 and 1239 unique department codes, and 4,777,003 and 2,049,079 contacts, respectively. We refer to these data sets as the *reduced DNPR2 training data* and *reduced DNPR2 validation data*, respectively.

DNPR3

Contacts starting between 1 February 2019 and 31 December 2020 were included to enable direct comparison with DNPR2. Non-physical contacts (eg virtual consultations) were excluded as DNPR2 only contained physical contacts. Contacts with negative duration or unknown department were also excluded.

The final DNPR3 data set contained 1660 unique department codes and 27,694,584 unique contacts. The DNPR3 data set was split into 70% training and 30% validation sets, with 1649 and 1632 unique department codes, and 19,385,948 and 8,308,636 contacts, respectively.

Covariates

The objective of this study was to classify the patient type of contacts based on the following variables, which are routinely registered by healthcare professionals and reported to DNPR:

- Contact duration: defined as time elapsed from the start timestamp to the end timestamp for a given contact. In DNPR2, duration is only available in whole hours.
- Contact cause: for example, disease, accident, etc.
- Admission type: whether a contact is acute or planned (elective).
- Overnight: defined as contacts that did not start and end on the same date (in Skjøth et al overnight contacts last from at least 5 pm to 7 am the next day).[9]
- In the supplementary material (section A.2) we provide a detailed description of the covariates.[11]

Statistical Methods

Tree-Based Classification

The objective of this study was to predict the patient type of contacts based on known variables, such as contact duration, contact cause, etc. To achieve this, we used tree-based classification methods, which are supervised learning algorithms commonly used in classification problems.[10] Further details on the classification models and software used are found in section A.3.1 in the supplementary material.

Clustering

Unlike supervised learning methods such as tree-based classification, which rely on labeled data for model training, unsupervised learning methods, such as clustering, identify patterns and insights from unlabeled data. This makes unsupervised learning suitable for uncovering patient types in DNPR3, where true labels are not available. Further details on the clustering methods and R packages used are found in section A.3.2 in the supplementary material.

While clustering is a powerful technique to uncover underlying structures within complex data sets, understanding the implications of each clustering can be challenging. To provide insights into the underlying patterns, we used classification trees to interpret the clustering results. By transforming clusters into classes, we revealed the defining features of each class.

Algorithm Performance

To measure the degree of agreement between the correct patient type labels and estimated labels, we used positive predictive values (PPVs) and sensitivities. Specifically, using the categorization of contacts registered in DNPR2 as the correct patient type (the gold standard), PPVs and sensitivities of the algorithm were computed for all hospital departments and their related hospital contacts in DNPR2.

In DNPR3, a similar approach was used, but with proxies for the true patient type labels. This is further described in the sections below.

To assess the degree of agreement between the proposed classification trees in DNPR3, we used the Adjusted Rand Index (ARI), which compares how closely two sets of estimated patient type labels match on a class-by-class basis, adjusting for chance. Note that $ARI = 1$ means perfect agreement while $ARI = 0$ means agreement by chance.

Department-Based Classification in DNPR2

One approach to classifying patient contacts involves considering the responsible department, as patient types are often concentrated in specific areas (eg acute outpatients are typically treated in emergency departments). An algorithm could first classify the department type, and this classification would then be applied to all contacts within that department. For example, if a department is classified as serving inpatients, all contacts in that department would be labeled as inpatient contacts.

We evaluated three department-based classification approaches using data from DNPR2: an approach similar to the classification algorithm proposed by Skjøth et al, a data-driven modification of this algorithm, and a best-case classification based on the most common patient type within each department.[9] The best-case approach served as a benchmark to assess the relative performance of the other two department-based approaches. The classification trees were built using the full DNPR2 training data and department labels and contact level comparisons to the true labels were done using the full DNPR2 validation data.

Department-Based Classification by Skjøth et al

Skjøth et al proposed a method for classifying hospital departments based on the proportion of overnight contacts and elective patients. Specifically, they proposed the following thresholds for their classification algorithm for departments:

- Inpatient: ≥50% overnight contacts,
- Elective outpatient: <50% overnight contacts and ≥50% elective patients,
- Emergency department: <50% overnight contacts and <50% elective patients.

The information needed for this classification is available in DNPR2 and DNPR3. When computing the proportion of overnight and elective contacts in DNPR2, outpatient visits were included as contacts. By convention, all elective outpatient visits (based on the true classification in DNPR2) counted as not overnight.

In Skjøth et al the performance of the classification algorithm was assessed by comparing its department classification with expert-provided labels, showing good concordance.[9] However, this does not indicate how well the algorithm performs at the individual contact level. In the present study, to evaluate this, we compared the classifications produced by the algorithm to the true patient type labels in DNPR2.

The error of a department-based classification approach, such as the one proposed by Skjøth et al, can be decomposed into two sources: error in the labeling of departments and variation in patient types within departments.[9] The first source of error occurs when the department-based algorithm does not label a department with its most common patient type, while the second occurs when a department contains multiple patient types, eg both acute outpatients and inpatients. We assessed the relative impact of these two sources of error by plotting the prevalence of the most common patient type (true DNPR2 labels) against the algorithm's department classification and the true most common patient type for each department. Further, the variation within departments was examined by considering a best-case scenario version of the department-based classification approach.

Data-Driven Department-Based Classification

To our knowledge, there was no empirical justification behind the classification thresholds used by Skjøth et al.[9] Therefore, we propose a data-driven modification using a supervised learning method, specifically tree-based classification. We refer to this method as the *Modified* department-based approach. The input variables were:

- Proportion of overnight contacts,
- Proportion of elective contacts.

The target feature was the true patient type as defined in DNPR2, with the categories: inpatient, elective outpatient, or acute outpatient.

Best-Case Department-Based Classification

To study the potential suitability of classifying patient types based on department types, we considered the performance of the best possible department-based classification algorithm. This algorithm assigns each department its (true) most common patient type. In other words, the first source of error of department-based approaches is eliminated leaving only the variation within departments as a source of error. While the true most common patient type for each department is not known in DNPR3, it could be approximated by manual expert classification.

We constructed this best-case classification algorithm by assigning each department in the full DNPR2 validation data to its most common patient type. Performance was evaluated in the full DNPR2 validation data ie, no training data were used to assess the best-case scenario.

Hybrid Department-Based Classification in DNPR2

In DNPR2, patient types were defined based on admission type, indicating whether the contact was elective or acute, see Table A.2 in the supplementary material. This approach divides outpatient contacts into two separate groups. However, using the department-based approaches described earlier, a contact that represents a minority within its department could be obviously misclassified based on the recorded admission type. For instance, an acute outpatient contact in a department with predominantly daytime elective contacts could be misclassified as an elective outpatient, clearly contradicting its actual admission type.

To address this issue, we expect that combining the admission type of contacts with the proportion of overnight contacts at the department level will improve accuracy compared to purely department-based approaches.

The optimal threshold for the department-level proportion of overnight contacts was determined by building two classification trees, one for acute contacts and one for elective contacts. Note that the proportion of overnight contacts was calculated by admission type.

Contact-Based Classification in DNPR2 and DNPR3

Department-based classification requires representative information on each department, such as the proportion of overnight contacts and elective patients, to determine the patient type for individual contacts. However, this may not always be feasible in practice.

To address this, we propose two classification algorithms based on individual contact data to identify the most likely patient type. We divided the classification problem into two binary classification problems: classification of elective contacts as either inpatient or elective outpatient, and classification of acute contacts as either inpatient or acute outpatient. From a clinical point of view, elective contacts that extend overnight should be classified as inpatient. Also, acute contacts with a duration $\geq$ 24 hours were classified as inpatient as only 0.5% of the acute outpatient contacts in the full DNPR2 data lasted longer than 24 hours. This forms the basis of the initial classification algorithm represented by the decision tree presented in Figure 1. Here, the solid arrows represent the predefined criteria, and the empty nodes and wavy arrows represent classification criteria to be learned. The input variables are listed below.

For acute contacts lasting <24 hours:

- Duration of contact (hours from start to end),
- Overnight stay (yes/no),
- Cause of contact (disease, accident, act of violence, attempted suicide, other intentional self-injury, injury reporting on subsequent contact, secondary contact after injury, other cause for contact, or unknown cause).

For elective contacts not extending overnight:

- Duration of contact (hours from start to end).

Note that the cause of contact is only mandatory to report for acute contacts.[12] Thus, essentially all elective contacts have an unknown cause of contact.

Since true patient type labels are unknown in DNPR3, two methods to address this were considered:

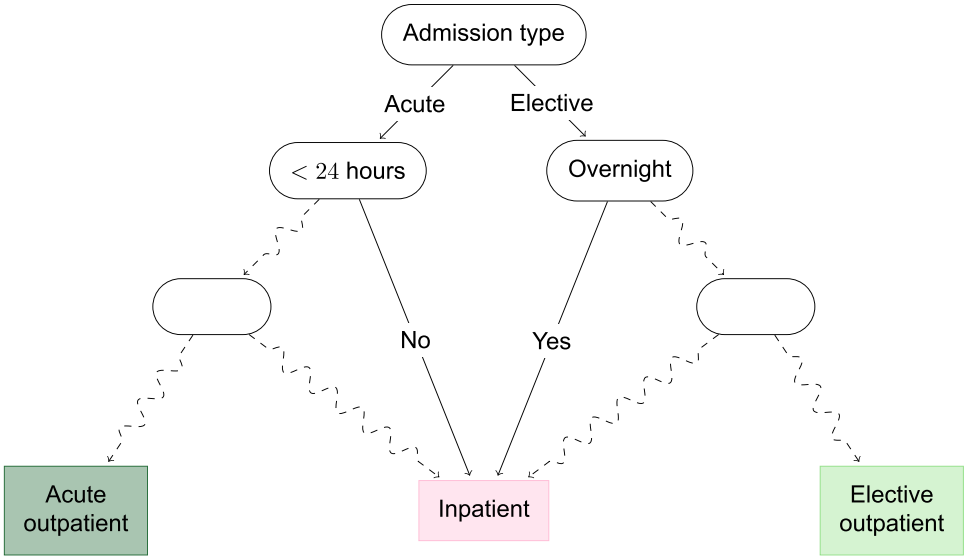

Figure 1 Decision tree skeleton. Solid lines represent a priori criteria used to determine a contact's patient type. Dashed lines represent the parts of the tree to be learned from data.

- Hybrid proxy labels: We used the labels estimated by the hybrid classification algorithm as proxies for the unobserved true patient types in DNPR3. This enabled supervised methods to learn the remaining classification criteria as represented by empty nodes and wavy arrows in the decision tree in Figure 1 using the hybrid proxy labels as responses.

- Cluster proxy labels: We employed a clustering method to learn clusters within acute contacts lasting <24 hours and non-overnight elective contacts. This enabled supervised methods to learn the remaining classification criteria as represented by empty nodes and wavy arrows in the decision tree in Figure 1 using the cluster proxy labels as responses.

These two proxy label approaches, detailed further in the sections below, were applied in both DNPR2 and DNPR3. The classification trees were built using the reduced DNPR2 training data and the DNPR3 training data. Additionally, a similar classification tree based on contact-level information was built using the true patient types in the DNPR2 training data as responses in order to assess the procedure in a setting where the true labels are known.

The concordance between the two proxy label approaches and their corresponding proxy labels was assessed using validation data. We report values similar to PPVs and sensitivities, treating the proxy labels as the "true" labels. We refer to these values as pseudo PPVs and pseudo sensitivities.

Classification Using Hybrid Proxy Labels

The hybrid classification algorithm was fitted on training data and the resulting classification of contacts as either acute outpatient, elective outpatient, or inpatient was used as proxy labels.

We then built a classification tree on the acute and elective branch separately, conforming to Figure 1, using the hybrid proxy labels as response. The resulting classification is referred to as the hybrid-based classification algorithm.

Classification Using Cluster Proxy Labels

We divided the clustering problem into two branches, conforming to Figure 1. For the acute branch, we used k-prototypes clustering to obtain two clusters representing inpatient contacts and acute outpatient contacts. For the elective branch, where the sole input variable was continuous (duration), we used k-means clustering with a pre-specification of two clusters representing inpatient contacts and elective outpatient contacts.

After clustering, we fitted classification trees to the acute and elective branches separately, using the estimated cluster memberships (cluster proxy labels) as response and the features described above. The resulting classification tree was examined, and each class was then manually labelled as either inpatient, elective outpatient, or acute outpatient based on the distribution of the characteristics as observed in DNPR2. The classification trees were built using training data and the resulting classification is referred to as the cluster-based classification algorithm.

Availability and Implementation

The proposed classification algorithms are implemented in R for direct application on DPNR3 data and available at: https://github.com/MARYs-DPH/DNPR

Ethical Considerations

The study is based solely on register data and thus, according to Danish law, no ethical approval is required and it is not required that the research be registered and approved by the Danish Data Protection Agency.[13] The data were only accessed in a pseudonymized version which is common practice when working on the Statistics Denmark server.[14]

Results

For the results of the exploratory data analysis, we refer to section A.4 in the supplementary material. For a summary and comparison of the results we refer to the final subsection of this section.

Department-Based Classification in DNPR2

Classification by Skjøth et al

We applied the algorithm proposed by Skjøth et al to the validation data, as described in the methods section.[9]

In Table 1 we present PPVs and sensitivities for the algorithm which compares the algorithm's classifications of contacts in the validation data with the true patient-type labels. The confusion matrix is found in Table A.3 in the supplementary material. The algorithm by Skjøth et al achieved excellent PPVs for inpatients and elective outpatients, though the PPV for acute outpatients was notably lower.[9] In contrast, the sensitivity for inpatients was moderate, while the sensitivities for elective and acute outpatients were excellent.

The two sources of error in department-based classification approaches, misclassification of departments and patient type variation within departments, were considered. In the validation data, the algorithm by Skjøth et al assigned 1644 departments to the label corresponding to the most common patient type in the department, leaving 82 departments incorrectly labeled.[9] One possible reason for this could be the fixed decision boundaries at 50% for the proportion of overnight contacts and elective patients within the department.

To explore this further, we reconstructed a figure in Skjøth et al illustrating the distribution of hospital unit classifications in DNPR2. However, we chose to color data based on the actual most common patient type (true labels from DNPR2), as shown in Figure 2.[9] Decision boundaries are indicated by dashed lines. The number of contacts per department varied substantially. The median number of contacts was 2938, the mean was 8301, and the maximum was 213,366. Many departments had few contacts, while a few had many. The size of the points in Figure 2 reflects the

Table 1 PPVs and sensitivities for the different department-based and hybrid approaches for contact classification, compared to the true labels from DNPR2. The performance metrics are based on the full validation data from DNPR2

Patient Type	PPV	Sensitivity
Skjøth et al		
Acute Outpatient	78.9	97.1
Elective Outpatient	99.1	99.6
Inpatient	99.3	76.5
Modified		
Acute Outpatient	95.6	94.1
Elective Outpatient	99.5	99.6
Inpatient	96.4	97.2
Best-case		
Acute Outpatient	93.7	96.9
Elective Outpatient	99.5	99.6
Inpatient	99.2	95.9
Hybrid		
Acute Outpatient	97.3	96.8
Elective Outpatient	99.8	99.9
Inpatient	96.7	96.7

Abbreviations: PPV, positive predictive values; DNPR2, Danish National Patient Register version 2.

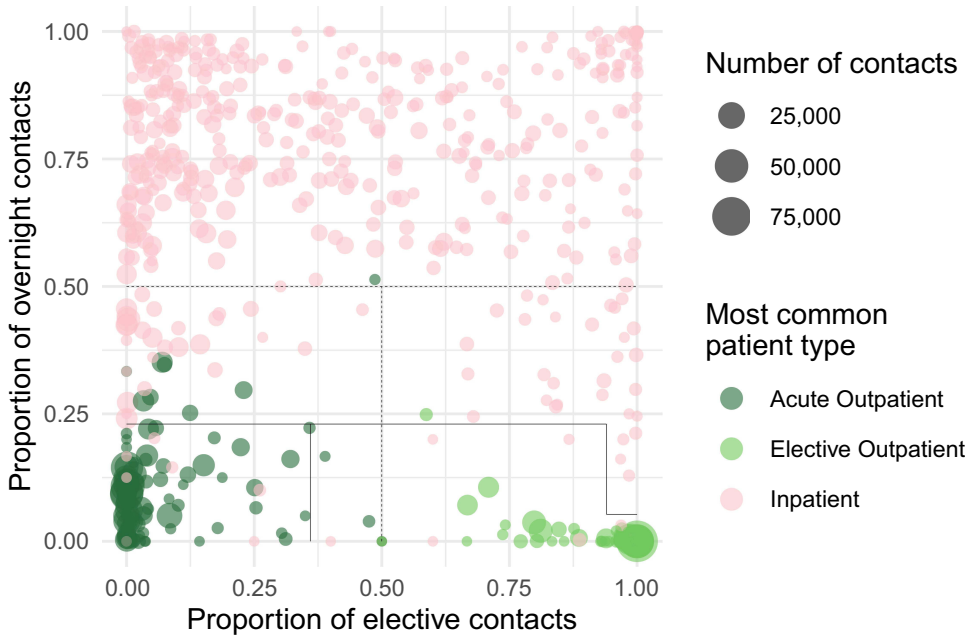

Figure 2 Distribution of hospital departments by the proportion of overnight and acute contacts, colored by the most prevalent patient type (based on DNPR2 labels) in the full DNPR2 validation data. Dashed lines represent decision boundaries suggested by Skjøth et al (2022), while solid lines represent the modified decision boundaries.

number of contacts per department. Visual inspection suggests that slight adjustments, such as lowering the threshold for overnight contact, could enhance performance. This is explored further in the next section.

The two sources of error were explored further in Figure 3, which shows the prevalence of the most common patient type (true DNPR2 label) for each department. We grouped the departments based on whether the algorithm's classification agreed with the most common patient type. The inpatient column *(c)* in Figure 3 shows that the departments

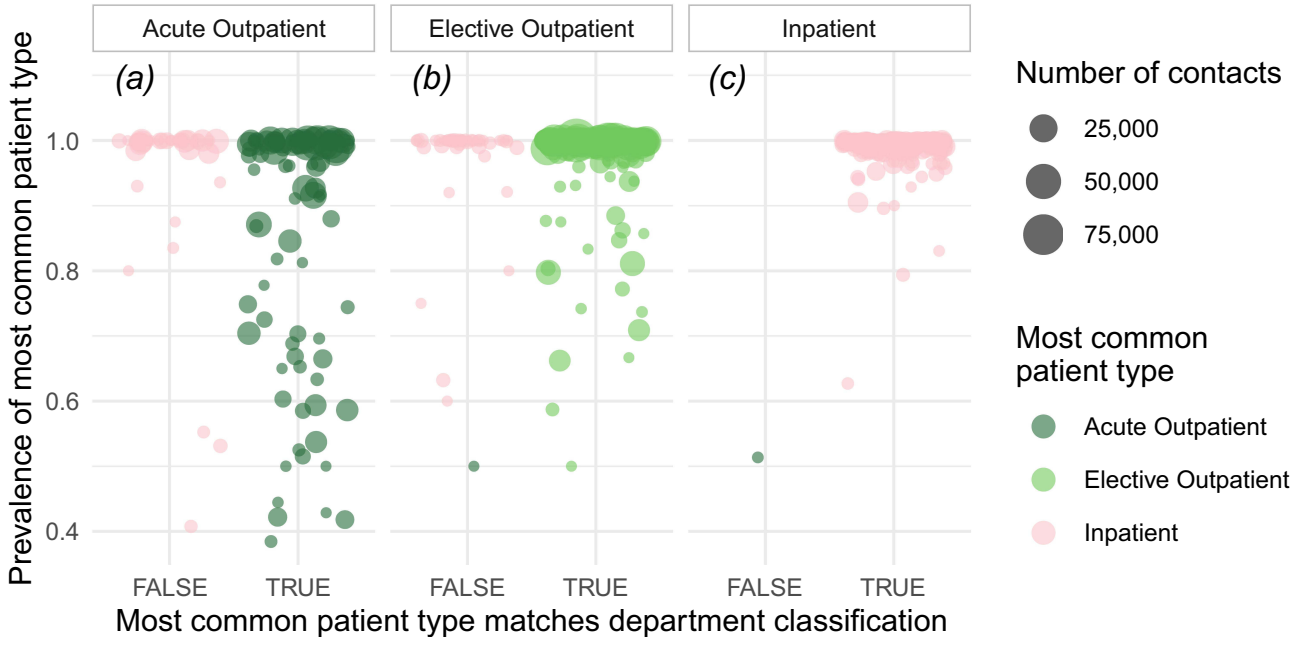

Figure 3 Prevalence of the most common patient type based on the DNPR2 labels acute outpatient (**a**), elective outpatient (**b**), and inpatient (**c**) at each department, classified by the algorithm proposed by Skjøth et al (2022), along with an indication of whether the algorithm correctly identified the department's most common patient type. The results are based on the full DNPR2 validation data.

https://doi.org/10.2147/CLEP.S500800

classified as inpatient primarily included inpatients (see for example the PPVs in Table 1), in fact, only one department was classified as inpatient despite having more contacts of another type (acute outpatient). However, as seen in columns *(a)* and *(b)*, many departments, and thus many contacts, were incorrectly not labeled as inpatient, resulting in lower sensitivity (Table 1). Departments classified as acute outpatient or elective outpatient by the algorithm mainly had the correct patient type, but some departments had mixed patient types or virtually no contacts from the category assigned by the algorithm.

Data-Driven Department-Based Classification

The following modified classification rules were found:

Inpatient

- If the department has ≥5% overnight contacts and ≥94% elective contacts *or*
- If the department has ≥23% overnight contacts and <94% elective contacts.

Elective Outpatient

- If the department has <5% overnight contacts and ≥94% elective patients *or*
- If the department has <23% overnight contacts and <94% and ≥36% elective patients.

Acute Outpatient

- If the department has <23% overnight contacts and <36% elective patients.

The classification algorithm is visualized in Figure 2, where solid lines represent the decision boundaries. The performance of this department-based algorithm for classifying contacts in the validation data is shown in Table 1. The confusion matrix is found in Table A.3 in the supplementary material. The modified algorithm showed a clear improvement in sensitivity for inpatients, and only a minor reduction in PPV. Furthermore, the PPV for acute outpatients clearly improved, with a minor reduction in sensitivity.

In total, the data-driven department-based algorithm classified 1691 departments in the validation data with the label corresponding to the most common patient type for the department, leaving 35 departments misclassified. This type of misclassification was twice as frequent in the original approach by Skjøth et al.[9]

Best-Case Department-Based Classification

Table 1 shows the performance of the best-case department-based classifier on classification of contacts in the validation data. The confusion matrix is found in Table A.3 in the supplementary material. This classifier performed very well, indicating that classifying contacts based on department type might be a valid option.

Hybrid Classification

The optimal thresholds were found to be 30% overnight for elective contacts and 23% for acute contacts resulting in the following classification rules for contacts:

Inpatient

- If the contact is elective and the department has ≥30% overnight contacts, *or*
- If the contact is acute and the department has ≥23% overnight contacts.

Elective Outpatient

If the contact is elective and the department has <30% overnight contacts.

Acute Outpatient

If the contact is acute and the department has <23% overnight contacts.

The performance of this hybrid classification algorithm on the full DNPR2 validation data is presented in Table 1. The confusion matrix is found in Table A.3 in the supplementary material. The hybrid approach demonstrated PPVs and

sensitivities comparable to the data-driven and best-case department-based approaches. Specifically, the hybrid approach improved both PPV and sensitivity for acute and elective outpatients, with only a slight reduction in sensitivity for inpatients compared to the data-driven department-based approach.

Data-Driven Contact-Based Classification in DNPR2

Based on the initial classification tree shown in Figure 1, we fitted classification trees on the reduced DNPR2 training data. As response, we used 1) the true patient type labels available in DNPR2, 2) labels produced by the data-driven hybrid classification algorithm as proxies for the true labels, and 3) labels produced by clustering as proxies for the true labels. The resulting classification algorithms are represented by the decision trees shown in Figures A.8, A.9, and A.10, respectively, in the supplementary material. In addition to the pre-specified variables (admission type, overnight, and whether the contact lasted more than 24 hours), the classification algorithms also use duration and cause of contact to classify contacts.

Table 2 compares the patient type labels produced by the three contact-based classification algorithms to the true patient types available in DNPR2. The confusion matrix is found in Table A.4 in the supplementary material.

When evaluating on the reduced DNPR2 validation data, we found that all three algorithms trained using true labels or hybrid proxy labels had excellent PPVs for elective outpatients, good PPVs for acute outpatients, and moderate PPVs for inpatients. On the other hand, they showed moderate performance regarding sensitivity for acute and elective outpatients and good performance in terms of sensitivity for inpatients.

The algorithm trained using cluster proxy labels showed excellent PPV for elective outpatients and good PPVs for acute outpatients and inpatients. On the other hand, it showed excellent sensitivity for acute outpatient, good sensitivity for elective outpatients, and moderate sensitivity for inpatients.

Table 2 PPVs and sensitivities of the classification algorithms using the true labels, using hybrid cluster labels, and using cluster proxy labels, relative to the true patient type. The performance is evaluated on the reduced DNPR2 validation data

Patient Type	PPV	Sensitivity
True labels		
Acute Outpatient	91.2	77.1
Elective Outpatient	99.3	76.5
Inpatient	72.9	93.1
Hybrid proxy labels		
Acute Outpatient	88.2	83.5
Elective Outpatient	99.3	76.5
Inpatient	75.6	89.9
Cluster proxy labels		
Acute Outpatient	77.7	96.9
Elective Outpatient	96.8	86.5
Inpatient	87.0	73.7

Abbreviations: PPV, positive predictive values; DNPR2, Danish National Patient Register version 2.

https://doi.org/10.2147/CLEP.S500800

Data-Driven Contact-Based Classification in DNPR3

Classification Using Hybrid Proxy Labels

Based on the initial classification tree shown in Figure 1 and using the patient type labels produced by the hybrid classification algorithm as proxies for the true labels, we fitted a classification tree on the DNPR3 training data. The resulting classification algorithm is represented by the decision tree in Figure 4.

In addition to the pre-specified variables admission type, and overnight, the decision tree also included contact duration to classify contacts.

Tables A.5 and A.6 compare the patient type labels produced by the hybrid approach with those derived from the decision tree in Figure 4. The values indicate a varying degree of agreement between the hybrid proxy labels and the classification. Specifically, identifying inpatients was difficult while identifying elective outpatients was more successful.

Classification Using Cluster Proxy Labels

The distribution of the contact duration within the resulting clusters for acute contacts, A and B, are visualized in Figure 5, and for the clusters for elective contacts, C and D, in Figure 6.

By comparing the variable distributions of the estimated clusters to the observed distributions in DNPR3 as well as the observed distribution within patient types in DNPR2, as shown in Figures A.2-A.7 in the supplementary material, cluster A was labeled as acute outpatient, clusters B and D were labeled as inpatient, and cluster C was labeled as elective outpatient.

We fitted a classification tree using the estimated cluster proxy labels as responses and with the same input variables used in the cluster analyses. The resulting classification algorithm is represented by the decision tree shown in Figure 7, and classifies contacts based on admission type, contact duration, and whether the contact was overnight.

Tables A.5 and A.6 compare the patient type labels produced by clustering in the DNPR3 validation data with those derived from the decision tree in Figure 7. These values indicate a very high degree of agreement between the hybrid proxy labels and the classification.

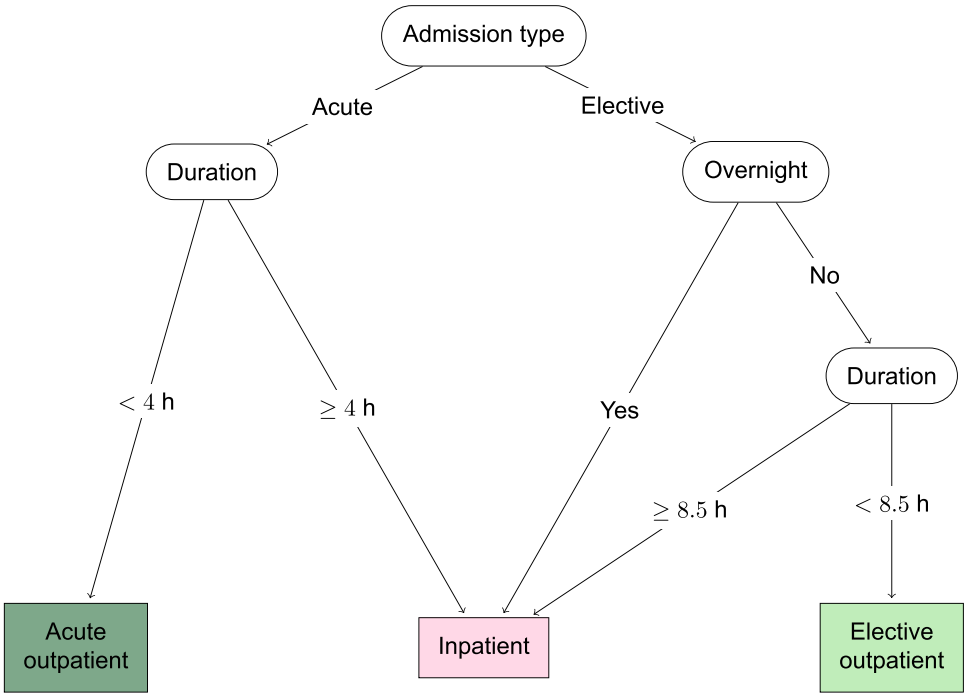

Figure 4 Illustration of the contact-based classification algorithm for DNPR3, using hybrid proxy labels, represented by a decision tree.

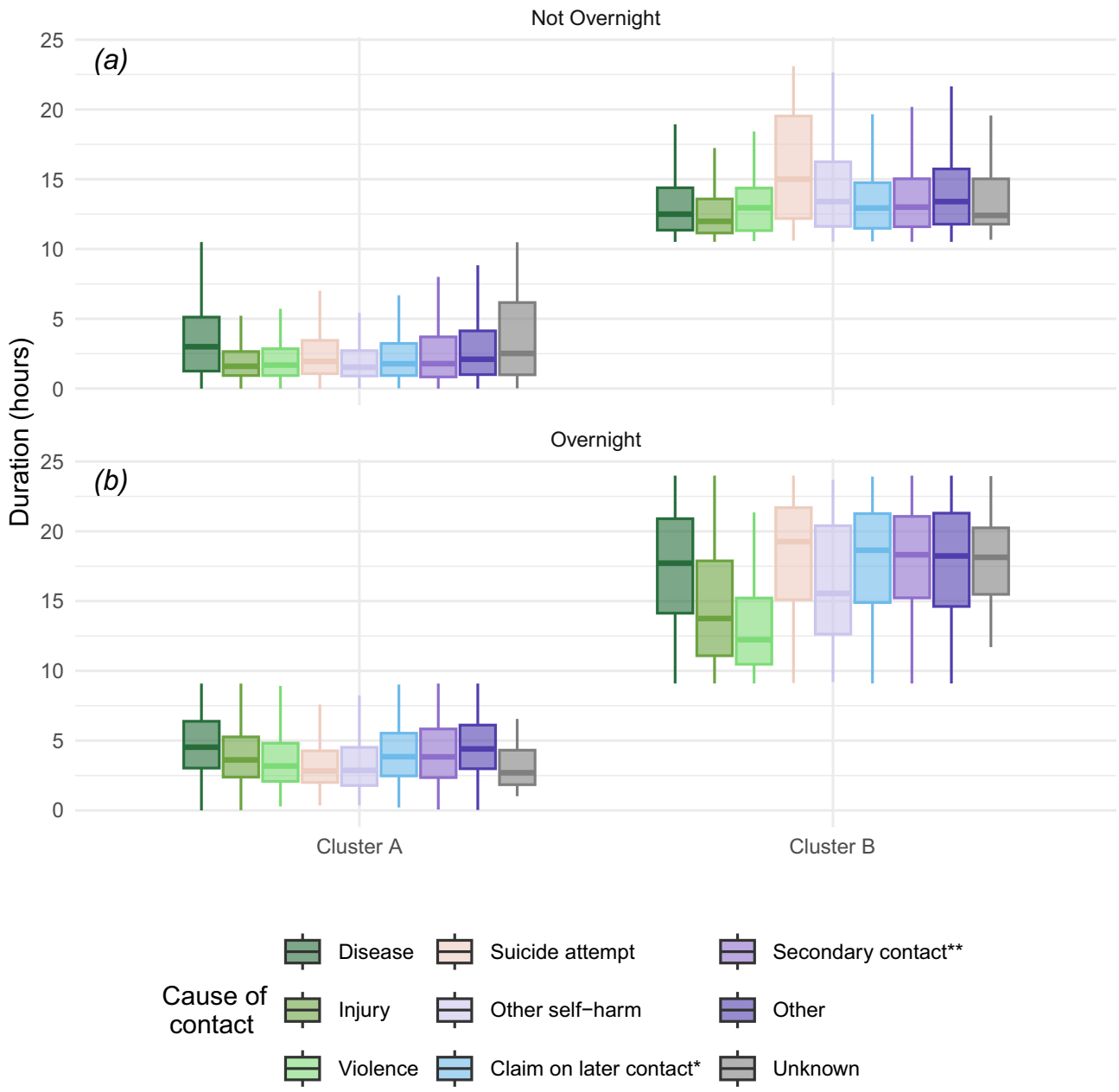

Figure 5 Boxplot of the duration (in hours) of acute contacts lasting < 24 hours and not extending overnight *(a)* or extending overnight *(b)*, grouped by estimated clusters and cause of contact in the DNPR3 training data. * Registration of claim will be completed on later contact. ** Secondary contact after claim.

Summary and Comparison of Classification Algorithms in DNPR3

We considered three different methods for classification in DNPR3: The hybrid approach combining proportion of overnight contacts at the responsible department with the admission type, the classification algorithm using contact level information seen in Figure 4 which was trained using the hybrid proxy labels, and the classification algorithm using contact level information seen in Figure 7 which was trained using the cluster proxy labels.

All three classification algorithms start by stratification based on the admission type of the contact, and subsequently classify contacts in the acute and elective branches as either in- or outpatient.

Compared to the classification algorithm using hybrid proxy labels, the classification algorithm using cluster proxy labels featured a higher decision boundary for distinguishing same-day elective outpatients from inpatients (8.5 hours

https://doi.org/10.2147/CLEP.S500800

Figure 6 Boxplot of the duration (in hours) of elective contacts that do not extend overnight, grouped by estimated clusters in the DNPR3 training data.

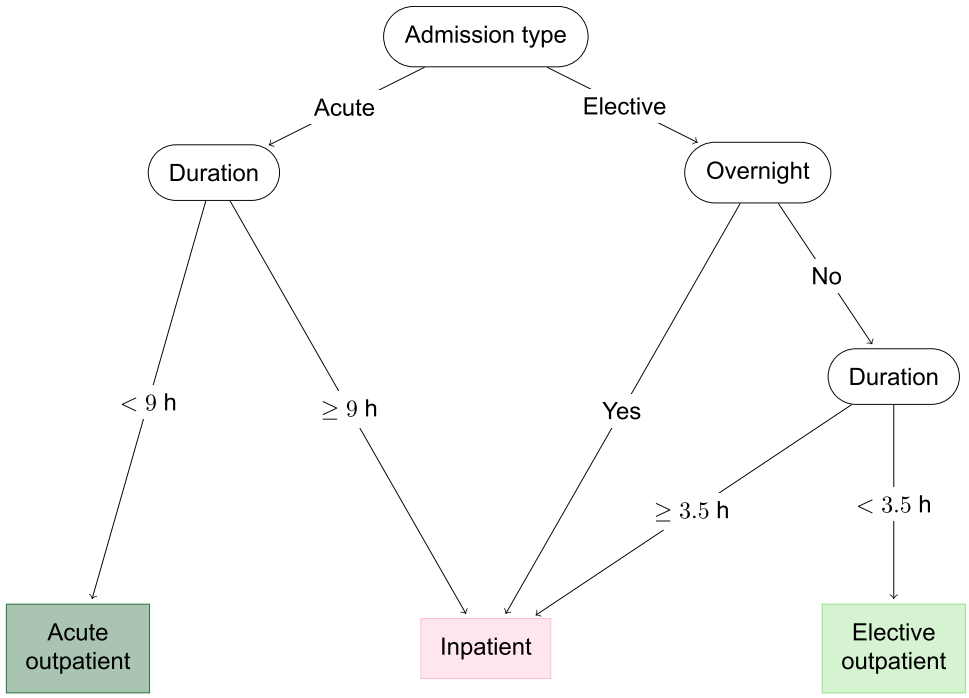

Figure 7 Illustration of the contact-based classification algorithm for DNPR3, using cluster proxy labels, represented by a decision tree.

compared to 3.5 hours). Conversely, the classification algorithm using cluster proxy labels featured a lower decision boundary for distinguishing acute outpatients from inpatients (4 hours compared to 9 hours).

Figure 8 shows the distribution of patient type classifications stratified by admission type in the DNPR3 validation data using the three classification algorithms. The classification algorithms produced comparable distributions of patient types. However, when using the two proxy label approaches, more elective contacts were categorized as inpatient compared to using the hybrid approach.

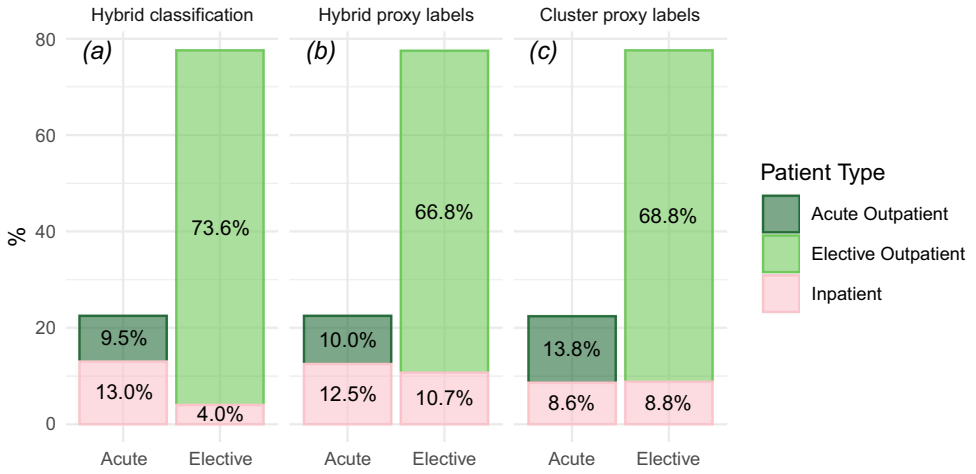

Figure 8 Distribution of estimated patient types in DNPR3 validation data, comparing the hybrid classification method (**a**) and the classification algorithms using hybrid proxy labels (**b**) and cluster proxy labels (**c**).

As a measure of the similarity between the classification algorithms we computed the ARIs. For the hybrid approach compared to using hybrid proxy labels (shown in Figure 4) and using cluster proxy labels (shown in Figure 7) we found ARI = 0.71 and ARI = 0.77, respectively. When comparing the proxy label approaches, we found ARI = 0.89.

This indicates a good match between the resulting algorithms for classifying patient types, even though they were derived through quite different methods.

Discussion

This study proposed and evaluated algorithms to classify hospital contacts in the third version of the Danish National Patient Register (DNPR) as inpatient, elective outpatient, or acute outpatient, aiming for a consensus on data interpretation among researchers in the field of register-based research in Denmark. Our goal was to ensure robustness and accuracy in our classification approaches by thorough exploration and application of various methods. Our study showed that all presented classification methods are suitable for categorizing patient types in DNPR2. The choice of method depends on the available data, ie, whether department-level or only contact-level data is available. The robustness of these methods supports their suitability in DNPR3. Specifically, it appears reasonable to start the classification by splitting contacts into acute and elective contacts, and subsequently classify contacts in these branches as in- or outpatients based on contact duration.

Overall, algorithms utilizing department-level information to classify individual contacts performed well in DNPR2. Notably, the best-case department-based classification achieved excellent sensitivity for all three patient types and excellent positive predictive value (PPV) for elective outpatients and inpatients. We revisited and modified an algorithm proposed by Skjøth et al, enhancing its performance through a data-driven adjustment of decision boundaries.[9] Additionally, a hybrid approach combining department-level and contact-level information demonstrated comparable PPV and sensitivity to the best-case department-based classification in DNPR2.

We introduced three methods for classifying contacts based on contact-level information: one trained in DNPR2 using true labels, one trained in DNPR3 through supervised learning with proxy labels based on the hybrid approach, and another trained in DNPR3 through unsupervised learning based on a clustering approach. These classification algorithms utilized variables such as admission type, cause of contact, overnight stays, and contact duration, as previously suggested by the Danish Health Data Authority.[8]

A notable strength of our proposed classification algorithms is their interpretability due to their tree structure, along with ease of implementation using the publicly available R code.

https://doi.org/10.2147/CLEP.S500800

Compared to prior studies that primarily relied on clinical considerations, our approach incorporated data-driven elements, leading to improved classification performance. We further advanced previous work by quantifying the performance of these methods for classifying contacts.

The department-based approach, as proposed by Skjøth et al and expanded upon in our study, shows promising results.[9] The variability within departments might be reduced by using the more detailed SOR classification in DNPR3, potentially enhancing the accuracy of the methods. However, this hypothesis remains untestable, as SOR information was not available in DNPR2.

In this paper, we focused on classifying contacts rather than departments, with department classification serving primarily as a steppingstone. The data-driven modification of the department-based approach was designed with this focus, assigning greater weight to those departments with a larger number of contacts.

The department-level classification approaches that performed well in DNPR2 are also applicable in DNPR3. However, these approaches require either representative data from departments to determine factors such as the proportion of overnight contacts or manual classification of departments. Given that the necessary data for the former might not be available to researchers due to limited study populations, and that the latter could be impractical given the large number of departments, applying a department-based approach may not always be feasible. Furthermore, even with suitable data, department-based classification algorithms can result in conceptually unreasonable classifications of contacts, such as labeling an acute contact as elective or labeling a week-long contact as outpatient. The hybrid approach addresses the admission type aspect of this problem, but not the duration aspect.

To overcome the limitations of department-level classification approaches, we proposed contact-based methods. These approaches only require data from the contacts of interest in DNPR3 to infer patient types, offering a clear advantage. Another benefit of the contact-based approaches is their transparency, as they provide an explicit categorization method for single contacts, enabling easy assessment of their plausibility based on clinical knowledge.

However, one main challenge in learning the optimal decision tree for classifying contacts by patient type remains: the lack of simultaneous availability of contact duration and true patient type in the same dataset. In DNPR2, the true labels are known, but the duration of elective outpatient visits was not recorded. Thus, training a supervised machine learning algorithm targeting the patient type in DNPR2 was only possible with elective outpatient records containing only one contact with plausible start and end times recorded. This imposed the untestable assumption that such contacts were representative of all elective outpatient contacts. Additionally, DNPR2 did not contain full minute-level data on the duration of records.

In DNPR3, the challenge arises from the absence of true patient type labels, which precludes direct training or evaluation of classification algorithms. The hybrid-based approach addresses this problem by using labels generated by the hybrid classification approach as proxies for the unknown true patient type. This approach allowed us to train a decision tree and evaluate its performance relative to the hybrid proxy labels, which exhibited good concordance. Given that the hybrid classification algorithm showed excellent performance in DNPR2, we anticipate similar performance in DNPR3; however, this is not verifiable. As an alternative, we considered an unsupervised approach using clustering of contacts in DNPR3. This approach accepts the absence of true patient type labels in DNPR3 and attempts to uncover underlying clusters of contacts. These clusters were subsequently labeled based on known characteristics of patient types. This presumes that the clusters reflect specific patient types. We evaluated the agreement between the methods using adjusted Rand Index (ARI) and while these values cannot be used to validate the methods, the high ARI values do indicate robustness of the methods.

Future research could explore more advanced classification techniques and compile comprehensive patient type information for all departments in Denmark to further enhance classification accuracy and applicability. Utilizing additional variables, such as patient age or diagnoses, could further enhance model accuracy.

While our study prioritized the interpretability in classification rules, introducing complexity to the algorithms might yield improved performance. Moreover, to address the limitations of the best-case department-based approach, future work could focus on compiling complete lists of primary patient types for hospital departments in Denmark, represented by SHAK or SOR codes. This would offer a more practical and scalable solution for implementing department-based methods.

Finally, our study focused on contacts at public somatic hospitals and contacts at private hospitals and psychiatric hospitals were not included. Future research could explore the generalization of the study to such contacts.

Conclusion

In conclusion, our study demonstrates that both department-based and contact-based classification approaches can effectively classify patient types in the Danish National Patient Register (DNPR). The direct interpretation of the algorithms aids researchers within the field of register-based research in Denmark.

When selecting a patient type classification algorithm, researchers should consider the intended use of the patient type (eg as an outcome or for inclusion/exclusion-criteria) and whether the study includes data from both DNPR2 and DNPR3 or only DNPR3 data. Based on these considerations, the choice between a department-based or contact-based approach should be made.

For studies favoring a department-based approach, we recommend utilizing the hybrid classification. For studies preferring a contact-based approach, we recommend using the hybrid proxy label classification algorithm if the study includes data from both DNPR2 and DNPR3 (using the true labels in DNPR2), and we recommend the cluster proxy label approach if the study relies solely on DNPR3 data.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Schmidt M, Schmidt SAJ, Adelborg K, et al. The Danish health care system and epidemiological research: from health care contacts to database records. *Clin Epidemiol*. 2019;11:563–591. doi:10.2147/CLEP.S179083
2. Sørensen ST, Kristensen FP, Troelsen FS, Schmidt M, Sørensen HT. Health registries as research tools: a review of methodological key issues. *Dan Med J*. 2023;70:A12220796.
3. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish national patient registry: a review of content, data quality, and research potential. *Clin Epidemiol*. 2015;449. doi:10.2147/CLEP.S91125
4. Danish Health Data Authority. The National Patient Register has been modernized [Danish: Landspatientregisteret er blevet moderniseret]. 2021.
5. Zimakoff AC, Jensen A, Vittrup DM, et al. Measles, mumps, and rubella vaccine at age 6 months and hospitalisation for infection before age 12 months: randomised controlled trial. *BMJ*;2023. e072724. doi:10.1136/bmj-2022-072724
6. Kildegaard H, Lund LC, Højlund M, Stensballe LG, Pottegård A. Risk of adverse events after covid-19 in Danish children and adolescents and effectiveness of BNT162b2 in adolescents: cohort study. *BMJ*. 2022;e068898. doi:10.1136/bmj-2021-068898
7. Statistics Denmark. Hospital utilization: statistical summary [Danish: Sygehusbenyttelse: statistisk behandling]. 2024 https://www.dst.dk/da/Statistik/dokumentation/statistikdokumentation/sygehusbenyttelse/statistisk-behandling.
8. Danish Health Data Authority. Guide for LPR3_F [Danish: Vejledning til LPR3_F]. 2022.
9. Skjøth F, Nielsen H, Bodilsen J. Validity of algorithm for classification of in- and outpatient hospital contacts in the Danish National Patient Registry. *Clin Epidemiol*. 2022;14:1561–1570. doi:10.2147/CLEP.S380023
10. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. *Routledge*. 2017. doi:10.1201/9781315139470
11. Documentation [Danish: Dokumentation]. *Danish Health Data Authority*. 2024 https://www.esundhed.dk/.
12. Danish Health Data Authority. Reporting guidelines for the National Patient Register Version 4.0 [Danish: Indberetningsvejledning til Landspatientregisteret Version 4.0]. 2024 https://sundhedsdatastyrelsen.dk/media/15194/LPR_Indberetningsvejledning_v.4.0.pdf.
13. Dansk Lov. The act on research ethics review of health research projects, section 14.2. www.retsinformation.dk. 2011.
14. Statistics Denmark. Access to data. 2024 https://www.dst.dk/en/TilSalg/Forskningsservice/Dataadgang.