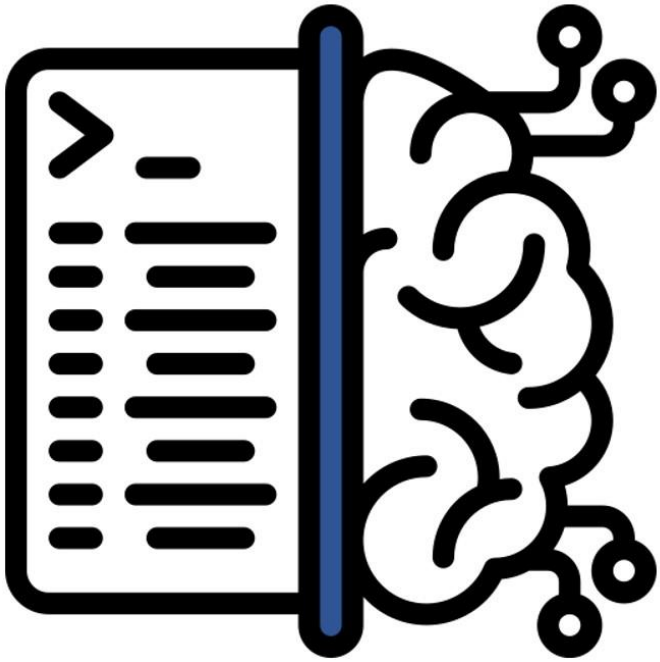
Improving Medical Care for Adults with Intellectual Disabilities (ID)

Can Automated Processing of Electronic Health Record Clinical Text Assist in ID Detection Among the General Outpatient Population?



Joyce Rijs

Master Thesis Technical Medicine

February 2024

TU Delft

Erasmus MC

Erfelijke & Aangeboren Aandoeningen

Inwendige Geneeskunde
Erasmus MC Rotterdam

This page was intentionally left blank.

Improving Medical Care for Adults with Intellectual Disabilities (ID): Can Automated Processing of Electronic Health Record Clinical Text Assist in ID Detection Among the General Outpatient Population?

Joyce Rijs

Student number: 4650611

February 27, 2024

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Dept. of Internal Medicine, Center for Adults with Complex Rare Genetic Syndromes

Erasmus MC

September 2023 – March 2024

Supervisors:

Dr. Laura C.G. de Graaff

Dr. Jifke F. Veenland

Thesis committee members:

Dr. Laura C.G. de Graaff, Erasmus MC (Chair)

Dr. Jifke F. Veenland, Erasmus MC

Dr. Renate G. Klaassen, TU Delft

Universiteit Leiden

TU Delft Delft University of Technology

Erasmus Universiteit Rotterdam

This page was intentionally left blank.

Preface

This thesis marks the culmination of my Technical Medicine studies, representing a significant milestone in my life. I am very grateful for all the opportunities I have had to grow both personally and professionally during my time as a student. Throughout my studies and internships, I became increasingly intrigued by the potential of Artificial Intelligence (AI) in healthcare. During this thesis, I had the opportunity to broaden my knowledge by becoming acquainted with Natural Language Processing (NLP) techniques. Combined with the clear social relevance of this project, this thesis was a perfect fit for me.

I would like to express my gratitude to Laura and Jifke, who supervised me throughout my thesis. Laura, your passion for improving medical care for adults with rare genetic syndromes is very inspiring. Thank you for inviting me to return to your department to perform my thesis and thereby contribute to your mission of improving ID detection. Jifke, thank you for introducing me to the world of Machine Learning (ML) and providing a critical perspective on my project. Thank you both for making time for me in your busy schedules, you motivated me to persevere through challenges and to think outside the box.

Furthermore, I want to express my appreciation to Jet, Michiel, Ruud, Jacquelien and Yon from Novicare for their enthusiasm for this project and offering me a warm welcome in Best. A special thanks to Jacquelien and Yon for their efforts in collecting data. Collecting data came with some unforeseen challenges, but I am proud that we were able to make the best of it.

Additionally, I am grateful to Tom Seinen for his invaluable assistance with the NLP techniques employed in this research.

To my fellow students and colleagues at the Center for Adults with Complex Rare Genetic Syndromes, thank you for your support, clinical guidance and enjoyable coffee and lunch breaks.

I also wish to thank Renate Klaassen for being part of my thesis committee.

Moreover, I extend my gratitude to my family, friends and roommates for their support during my studies. And, last but not least, I want to express my gratitude to Dave for always being there, listening to me, and supporting me throughout all my years of study.

Thank you all for being part of this journey! I am looking forward to putting everything I learned into practice in my future role as a data scientist at OLVG in Amsterdam.

Joyce Rijs
February 2024, Haarlem

Table of contents

List of abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
BMI	Body Mass Index
EHR	Electronic Health Record
GP	General Practitioner
ICPC	International Classification of Primary Care
ID	Intellectual Disability
IDA	Intellectual Disability Alert
ML	Machine Learning
NLP	Natural Language Processing
OCR	Optical Character Recognition
ROC	Receiver Operating Characteristic
SCAF	Screener for Adaptive Functioning
SCIL	Screener for Intelligence and Learning Disability
SOEP	Subjective, Objective, Evaluation, Plan
Std	Standard deviation
TF-IDF	Term Frequency – Inverse Document Frequency
UMLS-CUI	Unified Medical Language System Concept Unique Identifier

Summary

Background

Undetected Intellectual Disability (ID) can lead to chronic stress due to overestimation by society. Chronic stress can cause stress-related health issues, like hypertension, chronic fatigue and abdominal complaints. When a physician (General Practitioner (GP) or medical specialist) does not recognize that a patient has ID, the relation with stress may go unnoticed. In that case, the complaint is often treated as a purely somatic problem, while the underlying cause (overestimation due to unrecognized ID) remains untreated. This can increase healthcare consumption and impair the patient's quality of life. While physicians with ID-expertise can recognize subtle signs of mild ID, physicians without extensive experience will easily overlook the ID. To improve medical care for patients with ID, we aim to improve ID detection among physicians. As it is not feasible to give all individual doctors an 'ID-recognition training', we study the possibility of using AI to improve ID detection. In the past years, we have been working on an 'ID Alert' (IDA) using ML. In previous phases of the IDA project, structured Electronic Health Record (EHR) data was used for the creation of an IDA. In addition, in the current study, we investigate the use of unstructured EHR data (clinical text).

Methods

We analyzed unstructured correspondence files of 200 ID-adults and 200 non-ID adults of Novicare, an organization that provides multidisciplinary care to clients with complex and chronic conditions in intra- and extramural settings. Structured clinical data was unavailable. Therefore, we used an automated method of text extraction, de-identification and two types of feature extraction (bag-of-words and clinical concept extraction). Features were compared between ID-adults and non-ID adults. Significant features that were unlikely to be intrinsically different between ID- and non-ID adults were excluded. The remaining significant features were used for the training and evaluation (10-fold stratified cross-validation) of two Gradient Boosting Classifiers.

Results

Most features differed significantly between ID- and non-ID adults due to confounders such as differences in age, type of care and the doctor's word choice (which is inherent to the specialty and training of the doctor). Significant 'unbiased' features identified by both types of feature extraction methods are epilepsy, emotional disturbance (tension, arousal, agitation), visual or hearing problems and the presence of family members during consult. The developed ML models showed Areas Under the Curve (AUCs) of 0.98 and 0.89 for bag-of-words and clinical concepts, respectively.

Conclusion

This is the first study to investigate the use of unstructured correspondence files for developing an IDA. The developed models show a very high performance. Despite efforts to mitigate the effect of confounders, limitations may have influenced generalizability. Therefore, external validation of the proposed methods is necessary in future research.

Background

Individuals with ID have deficits in cognitive and adaptive skills necessary for daily societal functioning (1). ID can be classified into four levels of severity: mild, moderate, severe and profound (2). A Dutch study estimated the prevalence of all diagnosed subtypes of ID at 1.45% based on public service use, with a prevalence of 0.53% for mild ID exclusively (3). Cognitive impairment can be clear in some cases. However, ID can easily be missed in case of mild cognitive impairments, possibly in combination with relatively strong verbal skills.

Individuals with (undetected) mild ID are likely to be overestimated by society (4,5), which can cause stress and stress-related health issues, like hypertension, chronic fatigue and abdominal complaints (6–8). When a physician does not recognize that a patient has ID, the relation with stress may go unnoticed. In that case, the complaint is often treated as a purely somatic problem. This potentially results in unnecessary medication prescriptions, diagnostic procedures and medical complications (9,10), while the underlying issue (overestimation due to unrecognized ID) remains untreated. This can increase healthcare consumption and impair the patient's quality of life.

Research has shown that 42% of GPs in the Netherlands encounter challenges in recognizing mild ID (11). While physicians experienced in treating ID-adults can detect subtle signs of cognitive impairment based on years of experience, those without extensive experience will easily overlook the ID. Various societal, behavioral and clinical features (in the EHR) of a patient can trigger experienced physicians to consider testing for ID. However, it is not feasible to give all individual doctors an 'ID-recognition training'. Therefore, we aim to improve ID recognition among the general outpatient population by creating an automatic IDA with help from ML.

The IDA could warn physicians when an individual has a combination of characteristics that might be indicative of ID. Examples of health characteristics that differ between adults with and without ID include height, BMI, number and type of health problems and number and type of medication (12). The IDA does not replace the current diagnostic process, but triggers physicians of the possibility that an individual may have ID. When the IDA suspects ID, the next step is to confirm this suspicion with the Screener for Intelligence and Learning Disability (SCIL) and/or the Screener for Adaptive Functioning (SCAF), both pen-and-paper ID screener tools (13,14). Subsequently, the definitive ID diagnosis can only be made by detailed neurocognitive assessment. After diagnosis, adequate support can be arranged which will aid in daily societal functioning and reduce chronic stress (15). This support not only improves the quality of life but also reduces healthcare costs by avoiding unnecessary doctor's visits, diagnostic procedures and medication prescriptions. A recent publication in the Journal of Hypertension, about a 35-year-old female with hypertensive crisis whose blood pressure normalized after providing ID-specific support, stresses the importance of timely ID recognition (16).

At the Center for Adults with Complex Rare Genetic Syndromes of Erasmus MC, several technical medicine trainees have worked on the development of an IDA with different datasets from EHRs (unpublished work, (17–19)). During these studies, only structured EHR data was used. 'Structured data' refers to the information that is stored in distinct input fields, such as diagnosis codes and lab values. It is estimated that over 40% of EHR data contains unstructured information, which refers to clinical text (20). Using unstructured data for the development of ML models introduces new challenges in comparison to using structured data only, such as the need for de-identification and the extraction of useful information. Textual information is however not limited to code systems or distinct input fields and can therefore be more detailed than structured data. This study aims to build on previously developed IDAs (17–19) by utilizing unstructured EHR clinical data for the improvement of ID detection among the general outpatient population.

In this study, we collaborated with healthcare organization Novicare. Novicare provides multidisciplinary care to clients in their home environment and supports healthcare organizations in care for clients with ID and elderly clients. First, we explored the availability of both structured and unstructured client data for the development of a ML model. Subsequently, a method to automatically process the available data was proposed and used for the development of a ML model for the detection of ID.

Thesis structure

This thesis is organized into two parts. **Part A** focuses on data collection and serves as a preliminary study, laying the groundwork for the subsequent development of the methodology discussed in **Part B**, which focuses on data processing. The parts can be read separately. Goals, methods, results and discussion were presented separately for both **Parts A and B**.

Part A: Data collection (preliminary study)

Goals

This preliminary study focusses on data collection from Nova, Novicare's EHR system. The following research questions were answered in **Part A**:

- What client data is available in Nova?
- What improvements in data registration are necessary to obtain more reliable research data?
- Is it possible to collect enough reliable data for developing a ML model with the current data registration methods within Novicare?

Methods

Dataset and setting

Novicare clinicians report key clinical findings in Nova. Healthcare institutions where clients receive treatment often have a separate EHR system containing more elaborate client data. Medication information is stored in an external EHR system where both the healthcare institution and Novicare clinicians have access to.

In this study, we included two distinct client groups: 200 clients with ID ('ID-adults') and 200 elderly clients without cognitive impairment ('non-ID adults'). In order to match the age of the groups in the best way possible, we selected the 200 oldest ID-adults and the 200 youngest non-ID adults treated by Novicare. As no information about the level of ID was available, adults with all levels of ID were included in the ID group. Retrospective structured and unstructured client information in Nova was requested.

Statistical analyses

Client age and gender were statistically compared between the groups with a Student's t-test and a Chi-Square test (21), respectively.

Structured data

Data from all structured input fields in Nova was requested, including medication information, client appointments, diagnosis codes, lab values, information on intoxications (smoking, alcohol use, drugs), measurements (Body Mass Index (BMI), length, weight, blood pressure), and the occurrence of visual and hearing impairment.

Unstructured data

Two types of unstructured data were requested at Novicare. First, after receiving correspondence or conducting a consult, clinicians must report key findings in the EHR. These clinical notes are all written in the same structure with the SOEP-method (Subjective, Objective, Evaluation and Plan). All clinical notes were requested. Furthermore, all correspondence relating to a client, including GP files, questionnaires and all communication between healthcare providers was requested.

Recommendations

After thorough inspection of the available data, recommendations have been drawn up to improve the current data registration methods within Novicare to obtain more reliable research data.

Results

Dataset characteristics

Means and standard deviations (std) of age and gender of ID-adults and non-ID adults are shown in Table 1. Non-ID adults were significantly older than ID-adults ($p < 0.001$), while gender was comparable between the groups ($p = 0.10$).

	ID-adults (N=200)	**Non-ID adults (N=200)**	**P-value**
Age (mean ± std)	67 ± 7	78 ± 3	$p < 0.001$
Gender	45% male, 55% female	36% male, 64% female	$p = 0.10$

Table 1: Age and gender of ID-adults and non-ID adults.

Structured data

The availability of structured data is shown in Table 2. Medication information could not be obtained for privacy reasons. None of the structured data types received were known for all adults. Most structured data types showed higher availability for non-ID adults.

Type of structured data	**Availability for ID-adults (N=200)**	**Availability for non-ID adults (N=200)**
Medication	Not available	Not available
Client appointments	0%	87%
Diagnosis codes	23%	99%
Lab values	2%	75%
Intoxications (smoking, alcohol use, drugs)	0%	0%
Measurements (BMI, length, weight, blood pressure)	0%	0%
Visual impairment	1%	0%
Hearing impairment	0%	0%

Table 2: Availability of structured data for ID-adults and non-ID adults.

Unstructured data

The availability of unstructured data is shown in Table 3. Inspection of available clinical notes revealed that the clinical information in the notes of ID-adults was limited. Instead of reporting a summary of the most important findings, most clinicians referred to the correspondence for clinical information. The availability of correspondence was comparable between the groups and the available correspondence contained elaborate clinical information.

Type of unstructured data	**Availability for ID-adults (N = 200)**	**Availability for non-ID adults (N = 200)**
Clinical notes	72%	100%
Correspondence	93%	92%

Table 3: Availability of unstructured data for ID-adults and non-ID adults.

Recommendations

The recommendations for improvements of data registration at Novicare are shown in Appendix A.1. It is expected that if these recommendations are adopted within Novicare, the quality and quantity of client data will improve for future research.

Discussion and conclusion

In this preliminary study, the availability of structured and unstructured Novicare client data was explored. There is clearly a discrepancy in structured data registration methods between Novicare clinicians. Geriatricians fill in the structured input fields of non-ID adults more often than ID doctors for ID-adults. Some structured data is absent in both groups (intoxications, measurements, visual/hearing impairment). Possibly, this information is available at the healthcare institution where the client receives treatment and is not separately stored in Nova. Unfortunately, structured information from these healthcare institutions was not available for this project. The available structured data is insufficient for developing a ML model, especially in ID-adults where the information is too scarce.

The difference in data registration methods between both groups can also be seen in the clinical notes. Due to the limited clinical information in the clinical notes of ID-adults, the notes are also insufficient for developing a ML model. The recommendations on improvements of data registration might make the structured data and clinical notes more suitable for performing research in the future. Although these recommendations have been specifically addressed to Novicare, most recommendations might be generalizable to other healthcare organizations as well.

The limitations regarding the availability of structured data are not specific to Novicare. In previous research performed by our department with other datasets, the availability of structured information was also limited (17–19). During consult, the clinician takes the medical history, and documents relevant findings in clinical notes and structured fields. After taking the medical history, the clinician performs physical examination and reports findings in clinical notes and measurements in structured fields. However, some information that should be reported in structured fields, are documented as free text or left undocumented altogether. Various reasons for not completing all the structured fields may include time pressure, complexity of EHR interfaces or insufficient training (22).

Lastly, the availability of correspondence is comparable between both groups. Furthermore, the correspondence consists of useful clinical information in both groups. Therefore, we have decided to process the correspondence for the development of a ML model for the detection of ID. The difference in age between the groups will be carefully taken into account during the model development process.

Part B: Data processing (main study)

Goals

In **Part B**, the data collected in **Part A** was processed. After exploring available structured and unstructured client data at Novicare in **Part A**, we found that only client correspondence had sufficient clinical information to perform research. The following research questions were answered in **Part B**:

- Is it feasible to use NLP methods for unstructured clinical data in order to obtain ML features?
- Are there significant differences in the obtained features between ID-adults and non-ID adults?
- Is it possible to develop a ML model for the recognition of ID with at least comparable sensitivity and specificity as previous models developed by our department?
- How do the results compare with results from other datasets (previous research performed by our department)?

Methods

Dataset

In this study, we included two distinct client groups: 200 clients with ID ('ID-adults') and 200 elderly clients without cognitive impairment ('non-ID adults'). In order to match the age of the groups in the best way possible, we selected the 200 oldest ID-adults and the 200 youngest non-ID adults treated by Novicare. As no information about the level of ID was available, adults with all levels of ID were included in the ID group. All EHR correspondence files of the clients were used for analysis. Correspondence includes GP files, questionnaires and all communication between healthcare providers.

Statistical analyses

Client age and gender were statistically compared between the groups with a Student's t-test and a Chi-Square test, respectively. Furthermore, the amount of correspondence files per client and amount of words per client were statistically compared between the groups with a Mann-Whitney U test (21), as significant differences could introduce a bias.

Data processing

All data processing steps were performed in Python. Figure 1 gives an overview of the processing steps, including the development and evaluation of two ML models.

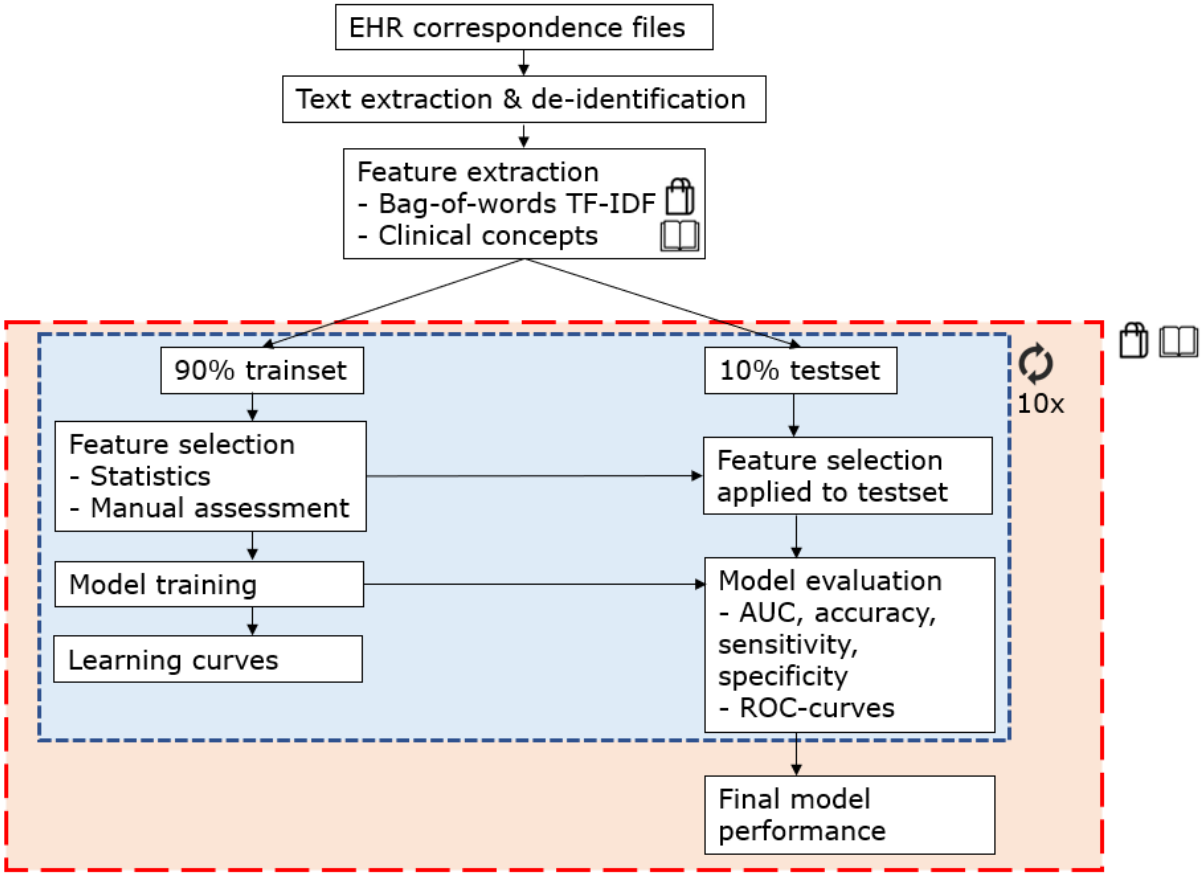


Figure 1: An overview of data processing.

De-identification

EHR correspondence contains sensitive information such as names and phone numbers that could identify clients. To ensure privacy and compliance with the European General Data Protection Regulation (23), a method to automatically de-identify correspondence was developed.

Text extraction for de-identification

First, the text from all correspondence files was extracted for de-identification. For every client, a folder with a zip-file containing all correspondence was provided. The zip-files were extracted, after which the text of every file was extracted with a method based on the document type. Text was extracted from the following file types: PDF, doc, docx, odt, txt, png, jpg, jpeg and tiff. For the extraction of text from PDFs and images, an Optical Character Recognition (OCR) tool called Pytesseract was used (24). For the extraction of text from doc files, Java was installed in order to use a Python package called Tika (25). The text of every file was saved in a txt file with the original extension of the file in the title and the number of the file. The text from a first PDF of a client was for example extracted and saved in a file called '1PDF.txt'. As the titles of the original files contained personal information, these were not kept. Furthermore, the file type (e.g. letter) was not systematically saved in the original title, meaning the file titles were of no use.

De-identification script

A script for automated text de-identification developed by L. van der Meulen (26) was modified and used for de-identification. Two types of replacements in text have been performed, one based on finding patterns (phone numbers, e-mail addresses, citizen service numbers and postal codes) and the other based on finding keywords (first and last names, institution names, countries, places, streets). Publicly available lists of the 10.000 most common first and last names in the Netherlands were utilized and supplemented with

all the client names at Novicare. Institution names that are clients of Novicare were also added to the lists. Furthermore, publicly available lists with countries, places and streets in the Netherlands were used.

If one of the regular expressions or keywords in the lists appeared in the text, it was replaced by a placeholder referring to the word type, such as <*NAAM*> (<NAME>) or <*PLAATS*> (<PLACE>). Keywords were only replaced if written with a capital letter. The de-identification process was approved by the privacy manager of Novicare and the client correspondence subsequently became available for use in the Erasmus MC.

Feature extraction

ML models cannot process raw text directly. To make the texts suitable for ML, two NLP feature extraction methods were employed: bag-of-words and clinical concept extraction. These methods were chosen because they are transparent and explainable, in contrast with more complex deep learning methods (27). The bag-of-words approach describes word occurrences in documents without considering context. Clinical concept extraction involves mapping texts to medical dictionaries for extracting medical information such as diseases, medications and medical procedures. The context of words is considered with this method. For feature extraction, all texts from the individual files were combined into one file for every client.

Bag-of-words

The text was tokenized into individual words ('tokens') by splitting the text at spaces and punctuation. Next, tokens that still contained any punctuation were split up ('symptoms/complaints' would for example be split up to tokens 'symptoms' and 'complaints'). Tokens only containing letters were included, meaning numbers and punctuation were excluded. Stopwords are common words that often do not carry significant meaning on their own and can occur frequently in a text (like 'the', 'is', 'and' etc.). In order to reduce dimensionality, common Dutch stopwords were removed from the list of tokens. The stopwords list was augmented with the placeholders used to mask personal information to ensure they were not included as tokens.

After tokenization, the binary occurrence of every token was counted for every client. Tokens that appeared in more than 80% or less than 20% of the clients were removed for dimensionality reduction, as these tokens were considered non-informative or too scarce. The remaining tokens were reduced to their base form with the Dutch version of Snowballstemmer (28) and the Term Frequency – Inverse Document Frequency (TF-IDF) of every stem was calculated with Formula S.1 (Appendix B.1). TF-IDF is a measure of how much information a word provides, adjusted for the fact that some words appear more frequently in general. Sci-kit learn (29) and NLTK (30) were used in Python.

Clinical concept extraction

MedSpacy, a clinical text processing toolkit enabling clinical concept extraction and context detection (31), was used for extracting clinical concepts. The English dependencies and language context rules were replaced by Dutch versions by T. Seinen et al. (32). Dutch context rules were applied to determine the context of the concepts, such as whether they were negated, mentioned in a hypothetical or historical context, or related to the client or someone else. Six Dutch vocabularies described in Table S.1 (Appendix B.2, adopted from (32)) were utilized to identify and categorize clinical concepts. The clinical concepts in all dictionaries are linked to standardized Unified Medical Language System Concept Unique Identifiers (UMLS-CUI), which were used for further analysis. After splitting the text into individual sentences, a word was mapped to a clinical concept if at least 70% of the word overlapped with the corresponding clinical concept. For example, the word 'unconscious' would be mapped to clinical concept 'unconsciousness', making the method more robust for word variants. Furthermore, this approach allowed for some leniency towards spelling

errors. Occurrence was calculated for every clinical concept. Additionally, the total number of diagnoses per client was calculated and included as a feature.

Train-test split

After feature extraction, the dataset was divided into a trainset containing 90% of the data and a testset containing 10% of the data in a stratified manner, meaning that the proportion of samples for both groups was preserved. To robustly evaluate performance and reduce the risk of overfitting (ensuring the model generalizes well to unseen data), a 10-fold cross-validation was performed. This involves defining and evaluating the model across ten different train-test splits.

Feature selection

For both TF-IDF features and clinical concepts, statistics were calculated in order to identify features that differed significantly between ID-adults and non-ID adults. A Mann-Whitney U test was performed for TF-IDF features and a Chi-square test was performed for clinical concepts (21). For the total amount of diagnoses per client, a Mann-Whitney U test was performed. All statistical analyses were performed in the trainsets, and a Holm-Bonferroni correction was applied to the p-values to account for multiple testing. This correction adjusts the significance levels based on the number of individual tests (33).

The features that were significantly different between ID- and non-ID adults in more than 5 of 10 cross-validation folds were evaluated for potential confounders, including differences in age, type of care and the doctor's word choice. It was essential to avoid incorporating features influenced by these confounders into the model, as this could lead to the model predicting age, type of care and the doctor's word choice rather than ID occurrence. Therefore, features that were thought to be unlikely intrinsically different between ID- and non-ID adults were excluded. J. Rijs, dr. L. de Graaff and dr. J. Veenland evaluated the features based on clinical experience and literature review. The remaining features were used for the development of two ML models.

Machine learning

Two Gradient Boosting classifiers were trained - one using TF-IDF features and one using clinical concepts. A Gradient Boosting classifier sequentially combines predictions of multiple decision trees (34). It minimizes errors from previous decision trees in order to enhance model performance. The model can capture complex relationships in data and can tolerate outliers and noise. The Sci-kit learn implementation of the Gradient Boosting Classifier was employed with the model parameters described in Table S.2 (Appendix B.3).

Model evaluation

For model evaluation, scoring metrics (mean AUC, accuracy, sensitivity and specificity) were computed across the test sets of the different folds. Receiver Operating Characteristic (ROC)-curves were generated, depicting sensitivity versus 1-specificity over different decision thresholds. The AUC of ROC-curves is a measure of accuracy. Sensitivity and specificity represent the conditional probability of correctly identifying true positives and true negatives, respectively. Accuracy represents the percentage of correctly predicted outcomes. Learning curves were created on the trainsets to visualize the behavior of the models as the amount of training data increases (35).

Results

Dataset characteristics

Table 4 shows the characteristics of the dataset. Non-ID adults were significantly older than ID-adults ($p < 0.001$), which could introduce bias in the results. This was taken into account in selecting features for model development. Gender, amount of files and amount of words were comparable between the groups ($p > 0.05$). 14 ID-adults and 17 non-ID adults had no EHR correspondence files.

	ID-adults (N = 200)	**Non-ID adults (N = 200)**	**P-value**
Age (mean ± std)	67 ± 7	78 ± 3	$p < 0.001$
Gender	45% male, 55% female	36% male, 64% female	$p = 0.10$
Missing data (amount of clients)	N = 14	N = 17	
Files per client (median)	5	6	$p = 0.45$
Words per client (median)	3821	4578	$p = 0.29$

Table 4: Dataset characteristics.

Data processing

Feature extraction and selection

Bag-of-words

After the calculation of bag-of-words tokens, TF-IDF differed significantly in 430 tokens. Most features significantly differed because of confounders instead of the difference in ID occurrence. Examples of biased words are 'verpleeghuis' (nursing home; more prevalent in non-ID adults due to age and type of care) and 'consultvoorbereiding' (consult preparation; after manually assessing the texts, it was identified that ID doctors often write a separate paragraph for preparation, whereas geriatricians do not). Only the 29 features described in Table S.3 (Appendix B.4) identified as (most likely) different due to ID occurrence were included for model creation. These are features about epilepsy, emotional disturbance, visual or hearing problems, communication, blood tests, family members (ID-adults are more likely to bring a family member to a consult) and others. These features were all more prevalent in the ID-adults, in accordance with clinical experience and/or literature (36–40).

Clinical concept extraction

After the extraction of clinical concepts, 237 features significantly differed between the groups. Again, most features probably differed due to confounding factors. After assessment for bias, 14 features described in Table S.4 (Appendix B.5) identified as (most likely) different due to ID occurrence were included for model creation. Features include epilepsy, emotional disturbance, autistic disorder, visual or hearing problems, operative surgical procedures, co-morbid conditions and family members. These features were more prevalent in the ID-adults, in accordance with clinical experience and/or literature (36–40). Figure 2 illustrates the features identified by bag-of-words, clinical concept extraction and both methods.

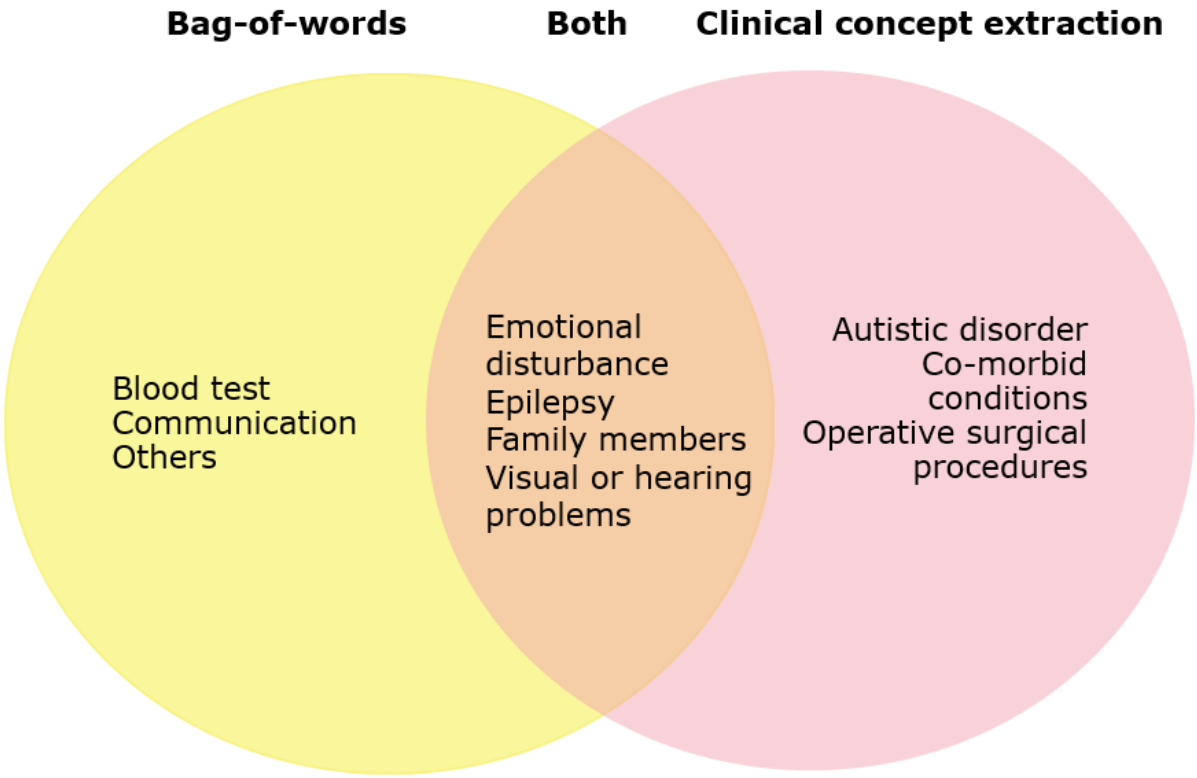


Figure 2: An overview of features more prevalent in ID-adults, identified by bag-of-words, clinical concept extraction and both methods.

Machine learning

After feature selection, two ML models were developed and evaluated. Table 5 shows the scoring metrics for the two models. The TF-IDF model has an AUC of 0.98 ± 0.01, an accuracy of 0.93 ± 0.04 and a sensitivity and specificity of 0.91 ± 0.06 and 0.95 ± 0.05 respectively. The clinical concepts model has an AUC of 0.89 ± 0.04, an accuracy of 0.83 ± 0.06 and a sensitivity and specificity of 0.84 ± 0.07 and 0.82 ± 0.11 respectively. The ROC-curves per fold are shown in Figures 3 and 4 for the TF-IDF model and the clinical concepts model, respectively. Learning curves (Figures S.1 and S.2 in Appendix B.6) show that the amount of data used for model development is sufficient, as the cross-validation score initially rises when more examples are added to the training set, and eventually stabilizes in both figures.

	TF-IDF	**Clinical concepts**
AUC	0.98 ± 0.01	0.89 ± 0.04
Accuracy	0.93 ± 0.04	0.83 ± 0.06
Sensitivity	0.91 ± 0.06	0.84 ± 0.07
Specificity	0.95 ± 0.05	0.82 ± 0.11

Table 5: Scoring metrics for the two models reported as mean ± std.

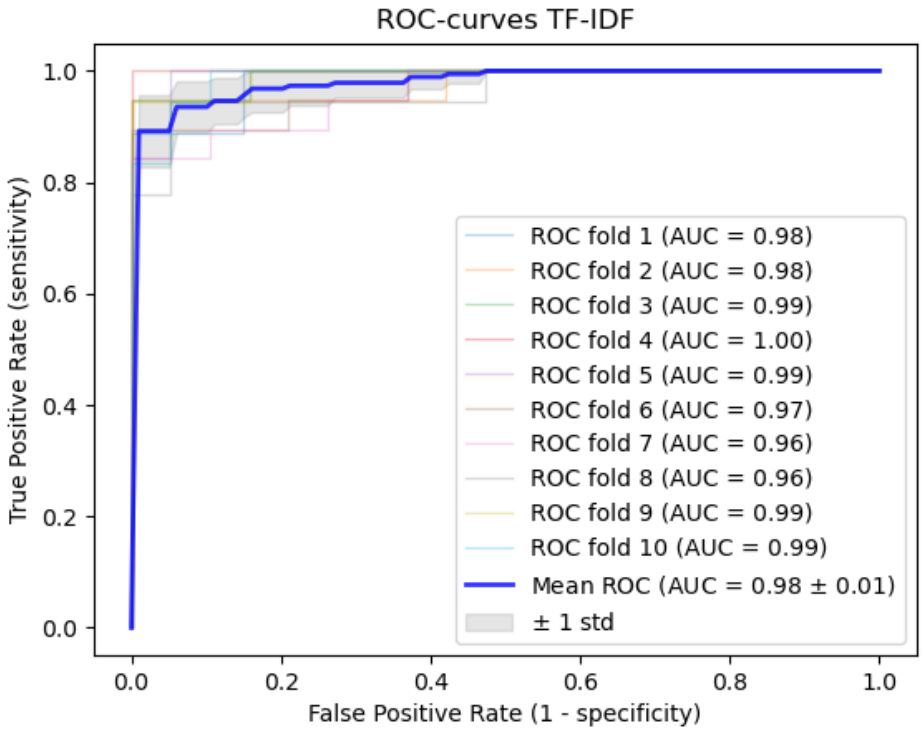


Figure 3: ROC-curves TF-IDF.

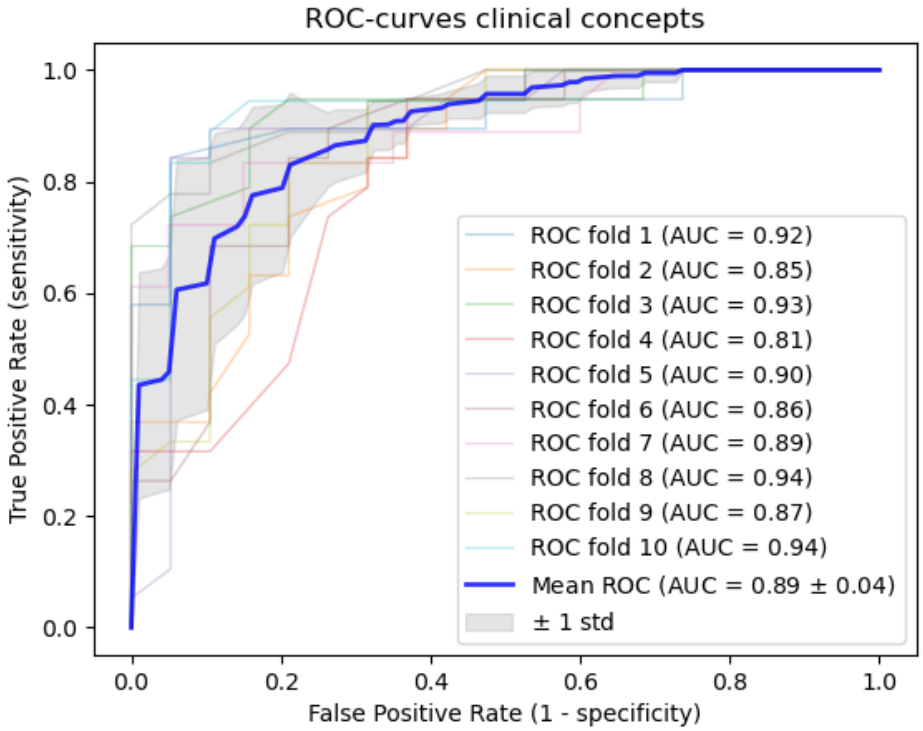


Figure 4: ROC-curves clinical concepts.

Discussion

In this study, a method to automatically process EHR correspondence for the development of two ML models for ID detection was developed. The automated processing consists of text extraction, de-identification and two types of NLP feature extraction methods. Features were compared between ID-adults and non-ID adults. Significant features that were unlikely to be intrinsically different between ID- and non-ID adults were excluded. The remaining significant features were used for the development of two Gradient Boosting classifiers. Significant 'unbiased' features identified by both types of feature extraction methods are epilepsy, emotional disturbance, visual or hearing problems and the presence of family members during consult. The developed models show high scoring metrics (AUC 0.98 and 0.89 for TF-IDF and clinical concepts, respectively). Despite efforts that were made to minimize the influence of confounders, limitations may still be present and have effects on the generalizability of the results.

The model created using the bag-of-words feature extraction method demonstrates an almost perfect mean AUC. However, bag-of-words features are sensitive to the specific word choice that characterizes the doctor's specialty, as synonyms are not mapped to the same feature. The majority of correspondence of ID-adults was written by ID doctors, whereas the majority of correspondence of non-ID adults was written by geriatricians. Most likely, these doctors use a different vocabulary, resulting in a confounder in model creation. Although efforts were made to exclude features influenced by confounders, the method was deemed insufficiently robust for the development of a generalizable model. Despite advantages in speed and transparency, the bag-of-words method is not recommended for datasets where texts from both groups are written by different types of doctors with distinct vocabularies.

Conversely, the clinical concept feature extraction method is less sensitive to differences in formulation, as synonyms are mapped to the same dictionary vocabulary. As a result, the outcomes of this method are considered more generalizable compared to the bag-of-words results, although confounders likely still influenced the outcomes. One of the confounders, the difference in type of care, is inherent to the dataset and could not be eliminated by excluding certain features. The features related to visual problems were for example picked as 'unbiased' features, as ID-adults more often experience visual problems. Possibly, ID doctors therefore pay specific attention to the presence of visual problems in their clients and describe them as such in their letters, while geriatricians do not specifically describe these problems in their correspondence. To validate the proposed methods, an external dataset with ID- and non-ID adults treated by the same type of doctor (such as GP or internist) is needed. For implementation of the method in the general outpatient population, it is important to note that both undetected ID- and non-ID adults receive the same type of care prior to ID diagnosis. This similarity in healthcare pathways helps to mitigate the impact of confounders such as different word choice and/or type of care. Therefore, NLP methods may be suitable for prospective data analysis in outpatient settings, where the influence of confounders is minimized.

Significant features identified by the bag-of-words and clinical concept extraction methods include epilepsy, emotional disturbance, visual or hearing problems and family members. The bag-of-words method also captured significant features related to communication, blood tests and other factors not captured by the clinical concept extraction method. Possibly, these words were different due to word choice and this bias was eliminated by clinical concept extraction. Conversely, the clinical concept extraction method captured autistic disorder, operative surgical procedures and co-morbid conditions, possibly due to its ability to map synonyms and identify clinical concepts more thoroughly.

In comparison with previous models developed with structured information (17–19), the models in this study demonstrate higher performance metrics (AUCs of 0.67-0.81 vs. AUCs of 0.89 and 0.98). However, it is important to approach the comparison of performances with caution, considering the bias present in the current project. Notably, certain significantly different features in this study align with those found in previous studies, including epilepsy, autistic disorder and co-morbid conditions. Figure 5 illustrates the features identified in previous research using structured data (17–19), features identified in this study using unstructured data, and features identified by both methods. In earlier research, different medication types were also identified as important model features. Unfortunately, structured medication information was unavailable in the current project. Medication features found with the clinical concept extraction method were not significantly more prevalent in ID-adults, possibly due to elderly individuals using more medication or ID doctors listing medication in correspondence less frequently than geriatricians.

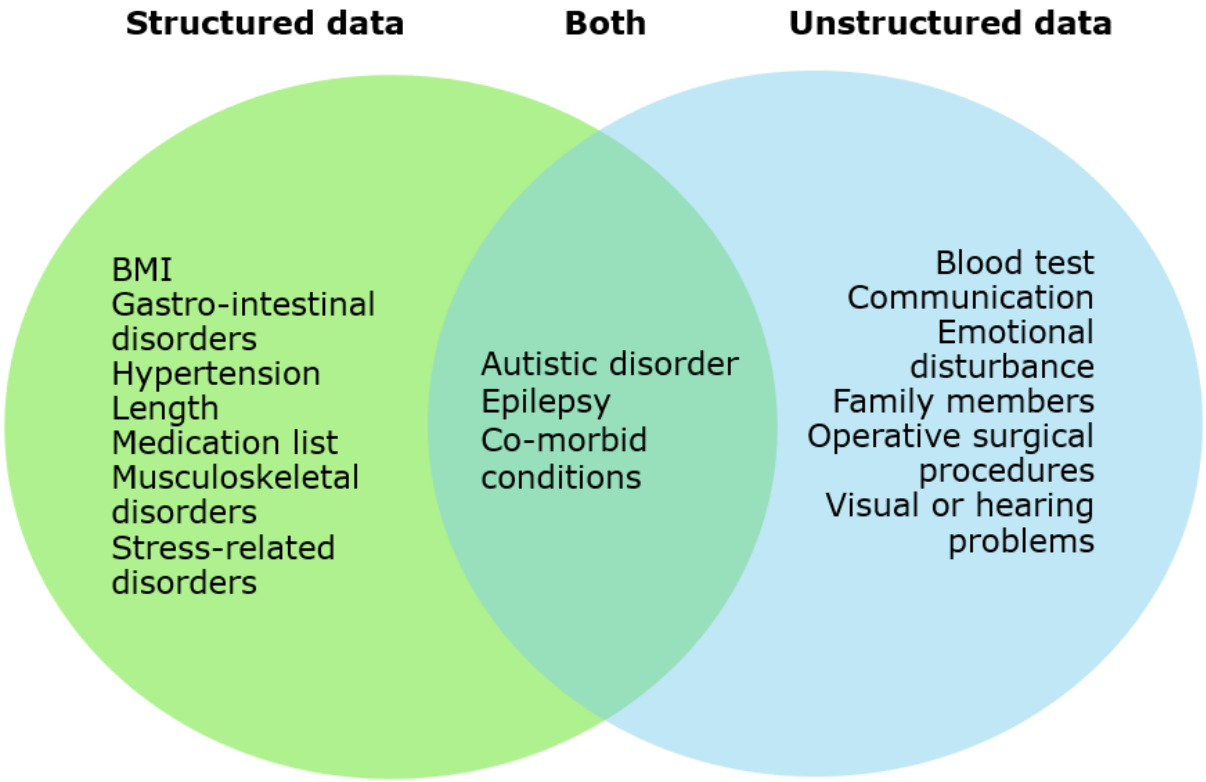


Figure 5: An overview of features that differ significantly between ID and non-ID adults identified in previous research using structured data, identified in this study using unstructured data, and identified by both methods.

Some features identified in the current project, such as emotional disturbance and visual or hearing problems, were absent in previous studies. This information is not stored in distinct input fields in the EHR, making it unavailable in datasets of structured information only. The absence of these features in structured datasets emphasizes the advantage of using unstructured information to enhance ID detection.

On the other hand, structured information contains unique information as well. Numerical structured data, such as measurements (BMI, length, weight, blood pressure) and lab values, were occasionally mentioned in correspondence. However, the current feature extraction methods only considered words. Structured and unstructured data have unique feature types that could complement each other when combined. The potential benefits of combining structured and unstructured data in prognostic prediction models have been

systematically reviewed by Seinen et al. (27). Using both types of data has been found beneficial in most studies. In future ID detection research, if both structured and unstructured information is available, exploring potential differences between data sources and investigating the combined value of both feature types for ML models would be recommended.

Limitations to this study include the absence of structured data, the lack of external validation, the relatively small sample size and the age difference between the groups. When defining the sample size, it was taken into account that the size should be large enough to create generalizable ML models, but small enough to get a reasonable match in age. The ML models were not further optimized in this study because of the assumed bias in the models. Additionally, the goal of the IDA is to promote earlier ID detection, especially in adults with mild ID. Since the type of ID was not available, adults with all levels of ID were included in this research. As a result, the findings may not be directly generalizable to the population with mild ID. Recommendations were made to Novicare to record the type of ID (**Part A**), allowing for more focused research on adults with mild ID in the future.

It should be acknowledged that some steps in the process were not thoroughly validated, such as the clinical concept extraction method. The clinical concepts attributed to clients should accurately represent the presence of the corresponding diagnoses or other clinical information, unlike the bag-of-words method, where the features simply represent word frequency. While the accuracy of the clinical concept extraction method was evaluated using a manually created test document, a more thorough validation of this method is recommended to ensure that correct clinical information is consistently attributed to clients.

When developing ML models for clinical use, it is important to address ethical concerns. A World Health Organization expert group has identified six key ethical principles guiding the development and use of AI in healthcare (41). These ethical principles have been carefully considered in this project. A main barrier to implementing ML in clinical practice is transparency (42). Therefore, relatively simple methods of de-identification and feature extraction have been chosen in order to promote transparency. It should however be acknowledged that the use of Gradient Boosting classifiers introduces some complexity in interpreting results. Future research should focus on balancing model performance with interpretability by comparing various classifiers.

Another key ethical principle is autonomy. It is important to emphasize that the IDA serves as a reminder and does not replace the current diagnostic process. Individuals retain autonomy, with the right to refuse participation in the diagnostic process if they wish. Additionally, the IDA promotes the ethical principle of equity by addressing the health disparities faced by ID-adults for various reasons (10). As part of the development of the IDA, health differences between adults with and without ID were explored (12). Enhanced knowledge of health issues in ID-adults could promote a more adequate approach to physical complaints. Furthermore, the clinical implementation of the IDA could contribute to earlier identification of ID, enabling ID-adults to receive appropriate assistance more promptly.

Conclusion

In conclusion, this is the first study to investigate the use of unstructured correspondence files for developing an IDA, a tool to increase ID-awareness among physicians. The definitive ID diagnosis can only be made by detailed and structured neurocognitive assessment, but the IDA can help physicians to recognize subtle signs of ID, which would lead to more timely and accurate use of ID-screeners like SCIL and SCAF.

In the current study, Gradient Boosting classifiers trained with features identified by the bag of words method and the clinical concept extraction method both showed a high performance in detecting ID. However, the model based on the bag of words features is very sensitive to word choice and therefore inherently biased. The model based on the clinical concept extraction features is less sensitive to the bias introduced by word choice, but further research is necessary to validate this method on an external dataset.

Significant 'unbiased' features identified by both types of feature extraction methods are epilepsy, emotional disturbance, visual or hearing problems and the presence of family members during consult. These features also emerged from previous phases of the IDA project or were confirmed by ID physicians as potentially relevant ID-predictors which, when occurring together in one individual patient, might be indicative of ID.

Combining structured and unstructured data in future research could enhance the development of a more accurate model promoting earlier ID detection, with the aim of improving quality of life for individuals with ID.

References

The icons used on the front page and in Figure 1 were derived from flaticon.com.

1. Patel DR, Apple R, Kanungo S, Akkal A. Narrative review of intellectual disability: definitions, evaluation and principles of treatment. Pediatric Medicine. 2018 Dec;1:11–11.
2. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th ed. Washington, DC: American Psychiatric Association; 2013. 33–40 p.
3. Cuypers M, Tobi H, Naaldenberg J, Leusink GL. Linking national public services data to estimate the prevalence of intellectual disabilities in The Netherlands: results from an explorative population-based study. Public Health. 2021 Jun 1;195:83–8.
4. Forte M, Jahoda A, Dagnan D. An anxious time? Exploring the nature of worries experienced by young people with a mild to moderate intellectual disability as they make the transition to adulthood. Br J Clin Psychol. 2011 Nov;50(4):398–411.
5. Snell ME, Luckasson R, Borthwick-Duffy WS, Bradley V, Buntinx WHE, Coulter DL, et al. Characteristics and needs of people with intellectual disability who have higher IQs. Intellect Dev Disabil. 2009 Jun;47(3):220–33.
6. Gawlik KS, Melnyk BM, Tan A. Associations Between Stress and Cardiovascular Disease Risk Factors Among Million Hearts Priority Populations. American Journal of Health Promotion. 2019 Sep 12;33(7):1063–6.
7. Konturek PC, Brzozowski T, Konturek SJ. Stress and the gut: pathophysiology, clinical consequences, diagnostic approach and treatment options. J Physiol Pharmacol. 2011 Dec;62(6):591–9.
8. Kocalevent RD, Hinz A, Brähler E, Klapp BF. Determinants of fatigue and stress. BMC Res Notes. 2011 Jul 20;4:238.
9. Ailey SH, Johnson TJ, Fogg L, Friese TR. Factors related to complications among adult patients with intellectual disabilities hospitalized at an academic medical center. Intellect Dev Disabil. 2015 Apr;53(2):114–9.
10. Krahn GL, Fox MH. Health disparities of adults with intellectual disabilities: what do we know? What do we do? J Appl Res Intellect Disabil. 2014 Sep;27(5):431–46.
11. Kramer P, Schalker M, ter Berg J. Kantar-rapport Herkenning van LVB door huisartsen. 2020 Apr.
12. Rosenberg AGW, Langendoen W, van der Lely AJ, Veenland JF, de Graaff LCG. Health differences between adults with and without intellectual disabilities at the internal medicine department: A first step to improve awareness of intellectual disabilities among healthcare professionals. Eur J Intern Med. 2022 Dec;106:154–7.
13. Nijman H, Kaal H, van Scheppingen L, Moonen X. Development and Testing of a Screener for Intelligence and Learning Disabilities (SCIL). J Appl Res Intellect Disabil. 2018 Jan;31(1):e59–67.
14. van Kessel S. Screener for Adaptive Functioning. [Amsterdam]: Universiteit van Amsterdam; 2017.
15. Scott HM, Havercamp SM. Mental health for people with intellectual disability: the impact of stress and social support. Am J Intellect Dev Disabil. 2014 Nov;119(6):552–64.
16. Bos-Roubos AG, Wingbermühle E, Giesen M, Kersseboom R, De Graaff LCG, Egger JIM. Hypertension with hidden causes: the cognitive and behavioral profile of an adult female with chronic stress and 16p11.2 microdeletion. J Hypertens. 2023 Sep 13;
17. Ruules L. "Intellectual Disability Alert" (IDA): a tool to improve medical care for adults with intellectual disabilities - CBS data. Rotterdam; 2023.

18. Chen L. v2.0 “Intellectual Disability Alert” (IDA): a tool to improve medical care for adults with intellectual disabilities - HiX data. Rotterdam; 2022.
19. Lindhout M. “Intellectual Disability Alert” (IDA): a tool to improve medical care for adults with intellectual disabilities - GGZ data. Rotterdam; 2023.
20. Dalianis H. Clinical text mining: Secondary use of electronic patient records. Clinical Text Mining: Secondary Use of Electronic Patient Records. Springer International Publishing; 2018. 1–181 p.
21. Conover WJ. Practical Nonparametric Statistics. 3rd edition. Wiley; 1999.
22. Bush RA, Kuelbs C, Ryu J, Jiang W, Chiang G. Structured Data Entry in the Electronic Medical Record: Perspectives of Pediatric Specialty Physicians and Surgeons. J Med Syst. 2017 May;41(5):75.
23. European Parliament, Council of the European Union. General Data Protection Regulation. 2016.
24. Hoffstaetter S, Bochi J, Lee M, Kistner L, Mitchell R, Cecchini E, et al. pytesseract 0.3.10 [Internet]. 2022 [cited 2024 Jan 30]. Available from: https://pypi.org/project/pytesseract/
25. Mattmann C. tika-python [Internet]. 2023 [cited 2024 Jan 30]. Available from: https://github.com/chrismattmann/tika-python
26. van der Meulen L. Towards Data Science. 2021 [cited 2024 Jan 30]. Remove personal information from a text with Python. Available from: https://towardsdatascience.com/remove-personal-information-from-text-with-python-232cb69cf074
27. Seinen TM, Fridgeirsson EA, Ioannou S, Jeannetot D, John LH, Kors JA, et al. Use of unstructured text in prognostic clinical prediction models: A systematic review. Vol. 29, Journal of the American Medical Informatics Association. Oxford University Press; 2022. p. 1292–302.
28. Snowball Developers. snowballstemmer 2.2.0 [Internet]. 2021 [cited 2024 Jan 30]. Available from: https://pypi.org/project/snowballstemmer/
29. Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python [Internet]. Vol. 12, Journal of Machine Learning Research. 2011. Available from: http://scikit-learn.sourceforge.net.
30. Bird S, Loper E, Klein E. Natural Language Processing with Python. O’Reilly Media Inc.; 2009.
31. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. AMIA Annu Symp Proc. 2021;2021:438–47.
32. Seinen TM, Kors JA, van Mulligen EM, Fridgeirsson E, Rijnbeek PR. The added value of text from Dutch general practitioner notes in predictive modeling. J Am Med Inform Assoc. 2023 Nov 17;30(12):1973–84.
33. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Vol. 6, Scandinavian Journal of Statistics. 1979.
34. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7:21.
35. Perlich C. Learning Curves in Machine Learning. In: Encyclopedia of Machine Learning. Boston, MA: Springer US; 2011. p. 577–80.
36. McMahon M, Hatton C. A comparison of the prevalence of health problems among adults with and without intellectual disability: A total administrative population study. Journal of Applied Research in Intellectual Disabilities. 2021 Jan 1;34(1):316–25.
37. Peklar J, Kos M, O’Dwyer M, McCarron M, McCallion P, Kenny RA, et al. Medication and supplement use in older people with and without intellectual disability: An observational, cross-sectional study. PLoS One. 2017 Sep 1;12(9).

38. Carey IM, Shah SM, Hosking FJ, DeWilde S, Harris T, Beighton C, et al. Health characteristics and consultation patterns of people with intellectual disability: A crosssectional database study in English general practice. British Journal of General Practice. 2016 Apr 1;66(645):e264–70.
39. Cooper SA, McLean G, Guthrie B, McConnachie A, Mercer S, Sullivan F, et al. Multiple physical and mental health comorbidity in adults with intellectual disabilities: Population-based cross-sectional analysis. BMC Fam Pract. 2015 Aug 27;16(1).
40. Smith M, Manduchi B, Burke É, Carroll R, McCallion P, McCarron M. Communication difficulties in adults with Intellectual Disability: Results from a national cross-sectional study. Res Dev Disabil. 2020 Feb;97:103557.
41. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva; 2021.
42. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform. 2021 Jan;113:103655.

Supplementary material

Appendix A.1: Recommendations for structuring Novicare data

These recommendations were written as a result of the data collection for the execution of the IDA project. These recommendations can assist in making client data more suitable for research in the future.

Structured data

- Consistently enter diagnoses for all clients. When treating a new client, store all useful information from the medical history in a structured manner in the EHR, including old lab results (structured).
- Save diagnoses in a coded manner as much as possible (as International Classification of Primary Care (ICPC) codes).
- Automatically save client appointments.
- Consider requesting known client characteristics (measurements such as BMI, length, weight, blood pressure) from the healthcare institution. This is valuable information for research.
- If known, record the type of ID a client has (mild, moderate, severe or profound).

Clinical notes

- Especially for ID doctors: provide a brief summary of the consult/letter; 'see correspondence' is too brief.

Correspondence

- File Titles:
    - Store titles in the same manner for all clients.
    - Include the file type (letter, dossier, correspondence) in the title with a clear abbreviation.
    - Avoid names, personal information, and client numbers in titles.
    - Example: 'Letter_AVG_20231026'.
- Avoid handwritten texts; type them out if possible, as it is easier to read and to recognize by the computer.
- Pay attention to the security of password-protected documents: only use passwords if absolutely necessary. Password-protected documents may be less usable or unusable for research.
- When scanning documents, ensure they are scanned in the correct orientation.

General recommendations

- Ensure that all care providers store information in the same way; this is useful for potential future studies.
- Avoid names in clinical notes and correspondence; use ‘Ms.’, ‘Mr.’ or ‘client’.
- Ensure that first and last names are written in capital letters (useful for automated anonymization).

Appendix B.1: Calculation of TD-IDF

$$Term\ Frequency\ (TF) = \frac{Number\ of\ times\ word\ appears\ in\ document}{Total\ number\ of\ words\ in\ document}$$

$$Inverse\ Document\ Frequency\ (IDF) = \log\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ where\ word\ is\ present}\right)$$

$$TF - IDF = TF * IDF$$

Formula S.1: Calculation of TF-IDF.

Appendix B.2: Dutch vocabularies used for clinical concept extraction

Abbreviation	Name	Source
SNOMED CT	SNOMED Clinical Terms and patient friendly terms	Dutch National IT Institute for Healthcare (NICTIZ)
MeSH	Medical Subject Headings	Unified Medical Language System (UMLS)
ICD10	International Classification of Diseases 10th revision	Unified Medical Language System (UMLS)
ICPC-1	International Classification of Primary Care	Unified Medical Language System (UMLS)
MedDRA	Medical Dictionary for Regulatory Activities	Unified Medical Language System (UMLS)
LOINC	Logical Observation Identifiers Names and Codes	Unified Medical Language System (UMLS)

Table S.1: Overview of Dutch vocabularies used as reference for clinical concept extraction.

Appendix B.3: Model parameters

Parameter	Used value
Loss	log_loss
Learning_rate	0.1
N_estimators	100
Criterion	friedman_mse
Min_samples_split	10
Min_samples_leaf	5
Max_depth	3

Table S.2: Model parameters used for the Gradient Boosting classifiers.

Appendix B.4: Selected TF-IDF features

Stem	Full words that were mapped to stem	Mean ± std ID-adults	Mean ± std non-ID adults	P-value[1]
emotionel	emotioneel, emotionele	0.03 ± 0.04	0.00 ± 0.01	p < 0.001
onrust	onrust, onrustig, onrustige, onrustigheid	0.03 ± 0.05	0.00 ± 0.01	p < 0.001
spanning	spanning, spanningen	0.02 ± 0.04	0.00 ± 0.01	p < 0.001
bos	boos, bos[2]	0.01 ± 0.02	0.01 ± 0.00	p < 0.001
rust	rust, rusten	0.01 ± 0.02	0.00 ± 0.01	p < 0.001
ontspann	ontspannen, ontspanning, ontspannend, ontspannende	0.01 ± 0.02	0.00 ± 0.00	p < 0.001
bril	bril	0.03 ± 0.06	0.00 ± 0.01	p < 0.001
ogen	ogen	0.03 ± 0.05	0.01 ± 0.01	p < 0.001
zien	zien	0.04 ± 0.04	0.01 ± 0.01	p < 0.001
draagt	draagt	0.02 ± 0.04	0.00 ± 0.03	p < 0.001
oren	oren	0.02 ± 0.05	0.00 ± 0.01	p < 0.001
hoort	hoort	0.01 ± 0.03	0.00 ± 0.00	p < 0.001
gehor	gehoor, gehorig, gehore	0.04 ± 0.04	0.00 ± 0.01	p < 0.001
prat	praten, praat, praatte	0.02 ± 0.02	0.00 ± 0.01	p < 0.001
duidelijk	duidelijk, duidelijke, duidelijkheid	0.03 ± 0.03	0.01 ± 0.01	p < 0.001
sprak	sprake, spraak, sprak, spraken	0.05 ± 0.05	0.02 ± 0.03	p < 0.001
epilepsie	epilepsie	0.04 ± 0.07	0.00 ± 0.02	p < 0.001
bloed-onderzoek	bloedonderzoek, bloedonderzoeken	0.01 ± 0.03	0.00 ± 0.01	p < 0.001
licham	lichamelijk, lichamelijke, lichaam, lichamelijkheid, lichamelijks	0.03 ± 0.03	0.02 ± 0.02	p < 0.001
geboort	geboorte	0.02 ± 0.03	0.00 ± 0.00	p < 0.001
vermoed	vermoeden, vermoedde, vermoedelijk, vermoedelijke, vermoed, vermoede, vermoedden	0.01 ± 0.02	0.00 ± 0.01	p < 0.001
dochter	dochter, dochters	0.03 ± 0.06	0.00 ± 0.02	p < 0.001
ouder	ouder, ouders, ouderen, ouderlijk, oudere, ouderlijk	0.03 ± 0.03	0.01 ± 0.02	p < 0.001
moeder	moeder, moederen, moederlijke, moeders	0.03 ± 0.05	0.00 ± 0.01	p < 0.001
zus	zus	0.04 ± 0.07	0.01 ± 0.03	p < 0.001
broer	broer, broers	0.03 ± 0.04	0.00 ± 0.01	p < 0.001
onderzocht	onderzocht, onderzochte, onderzochten	0.01 ± 0.02	0.00 ± 0.01	p < 0.001
afgenom	afgenomen	0.02 ± 0.04	0.00 ± 0.01	p < 0.001
vader	vader, vaders	0.02 ± 0.03	0.00 ± 0.00	p < 0.001

1. P-values were corrected with Holm-Bonferroni.
2. The Snowballstemmer does not include contextual word information, therefore bos was classified in the same stem as boos.

Table S.3: TF-IDF features that were significantly different between the groups and included for model creation.

Appendix B.5: Selected clinical concepts

UMLS-CUI	Definition	Mean ± std ID-adults	Mean ± std non-ID adults	P-value[1]
C0014544	Epilepsy	0.54 ± 0.50	0.12 ± 0.32	p < 0.001
C0543467	Operative surgical procedures	0.83 ± 0.37	0.50 ± 0.44	p < 0.001
C1275743	Co-morbid conditions	0.45 ± 0.50	0.12 ± 0.32	p < 0.001
C0030551	Parent[2]	0.60 ± 0.49	0.24 ± 0.43	p < 0.001
C0233494	Tension[3]	0.48 ± 0.50	0.17 ± 0.38	p < 0.001
C0085631	Agitation[3]	0.47 ± 0.50	0.17 ± 0.38	p < 0.001
C0337527	Brother[2]	0.41 ± 0.49	0.13 ± 0.34	p < 0.001
C0018767	Hearing	0.79 ± 0.41	0.50 ± 0.50	p < 0.001
C0004352	Autistic disorder	0.22 ± 0.41	0.02 ± 0.14	p < 0.001
C0337514	Sister[2]	0.52 ± 0.50	0.23 ± 0.43	p < 0.001
C0015421	Eyeglasses[4]	0.50 ± 0.50	0.21 ± 0.41	p < 0.001
C0042812	Visual acuity[4]	0.67 ± 0.47	0.37 ± 0.49	p < 0.001
C0026591	Mother[2]	0.51 ± 0.50	0.24 ± 0.43	p < 0.001
C0920139	Eyeglasses wearer[4]	0.26 ± 0.44	0.05 ± 0.22	p < 0.001

1: P-values were corrected with Holm-Bonferroni.
2: Features were combined into one ‘family’ feature.
3: Features were combined into one ‘emotional disturbance’ feature.
4: Features were combined into one ‘vision’ feature.

Table S.4: Clinical concept extraction features that were significantly different between the groups and included for model creation.

Appendix B.6: Learning curves

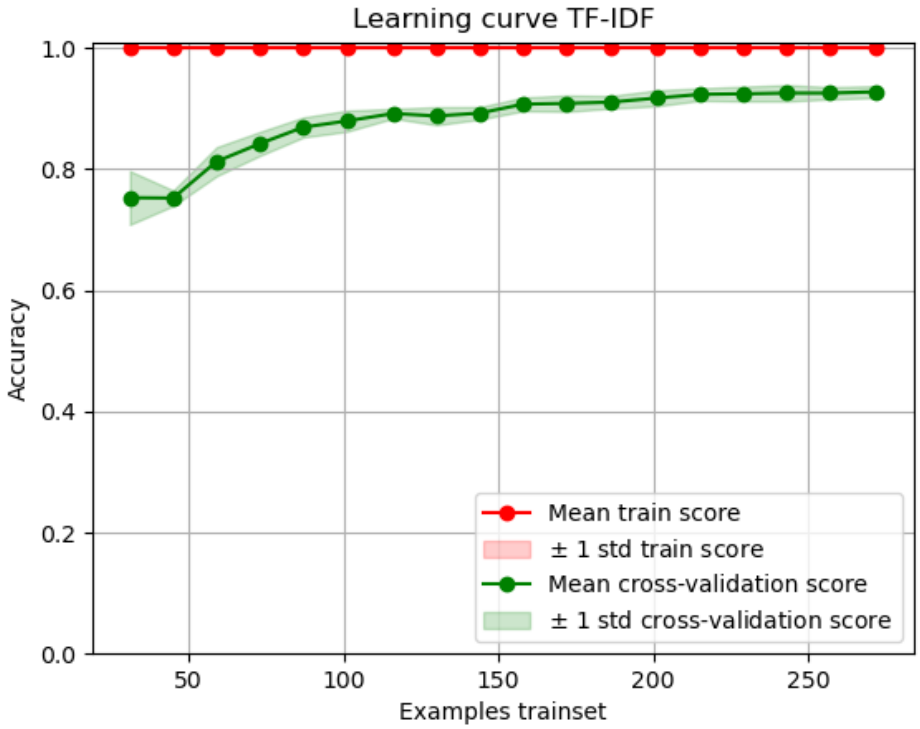


Figure S.1: Learning curve TF-IDF.

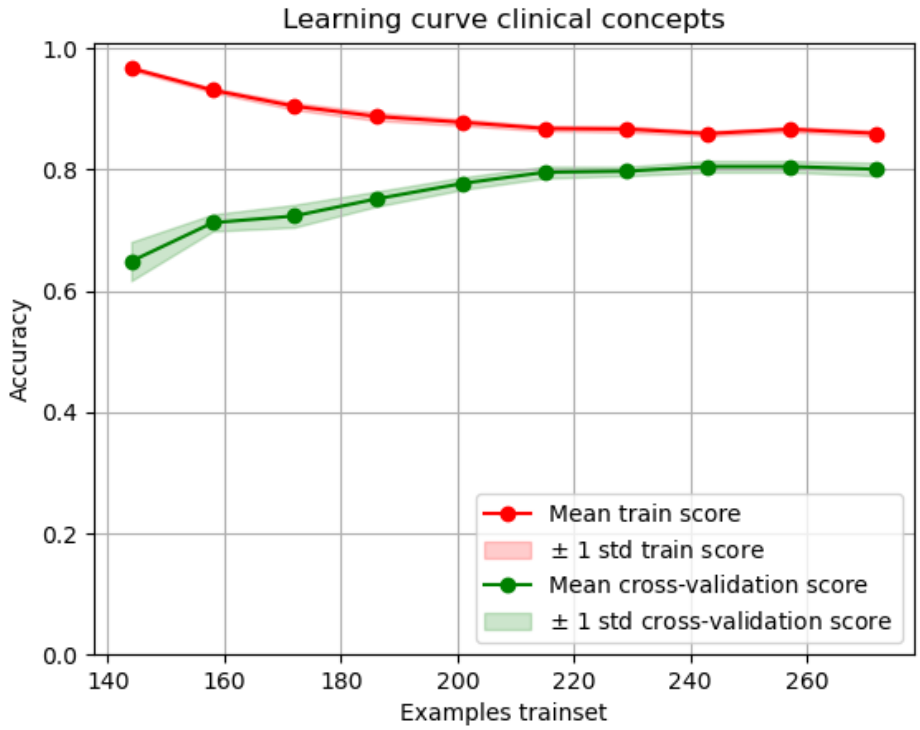


Figure S.2: Learning curve clinical concepts. This learning curve does not provide train and cross-validation scores for less than 144 examples in the trainset, therefore the x-axis is scaled differently than the x-axis for Figure S.1.