Clinical Information Extraction for Preterm Birth Risk Prediction

Lucas Sterckx[a,*], Gilles Vandewiele[a], Isabelle Dehaene[b], Olivier Janssens[a], Femke Ongenae[a], Femke De Backere[a], Filip De Turck[a], Kristien Roelens[b], Johan Decruyenaere[c], Sofie Van Hoecke[a], Thomas Demeester[a]

[a]*IDLab, Ghent University – imec*
Technologiepark-Zwijnaarde 126, Ghent, Belgium
[b]*Department of Gynaecology and Obstetrics, Ghent University Hospital*
Corneel Heymanslaan 10, Ghent, Belgium
[c]*Department of Intensive Care Medicine, Ghent University Hospital*
Corneel Heymanslaan 10, Ghent, Belgium

Abstract

This paper contributes to the pursuit of leveraging unstructured medical notes to structured clinical decision making. In particular, we present a pipeline for clinical information extraction from medical notes related to preterm birth, and discuss the main challenges as well as its potential for clinical practice. A large collection of medical notes, created by staff during hospitalizations of patients who were at risk of delivering preterm, was gathered and analyzed. Based on an annotated collection of notes, we trained and evaluated information extraction components to discover clinical entities such as symptoms, events, anatomical sites and procedures, as well as attributes linked to these clinical entities. In a retrospective study, we show that these are highly informative for clinical decision support models that are trained to predict whether delivery is likely to occur within specific time windows, in combination with structured information from electronic health records.

Keywords: Clinical Information Extraction, Clinical Decision Support Models, Preterm birth, Text mining

1. Introduction

In recent years, significant progress has been made in the area of machine learning for clinical decision support, mainly due to advancements of computational resources and the availability of electronic health record (EHR) data. Today, clinical decision support systems leverage EHR data to provide diagnostic and treatment recommendations at

*Corresponding author
Email address: lucas.sterckx@ugent.be (Lucas Sterckx)

the point of care, i.e., information personalized for the specific patient under consideration by the clinician at a given moment. This results in several advantages such as increased practitioner performance, improved quality of care, and better patient outcomes as the result of a more informed, evidence-based decision [1].

However, an estimated eighty percent of EHR data is composed of unstructured data or free-text notes compiled by doctors and nursing staff during patient encounters [2, 3]. Written language is a natural and expressive method to document clinical events and facilitate communication among the care team in the health care environment. The unstructured format of the notes permits recording of precise and domain-specific information for which the structured fields of the EHR may not be sufficiently detailed [4]. Clinical decision support systems preferably rely on structured data because of their transparency and interpretability for medical experts [1].

This paper studies the potential of automated information extraction (IE) from clinical text to enhance a decision support system. We intend to provide medical experts with background information to appreciate the merits and difficulties of clinical information extraction from unstructured texts. The released software, with suitable documentation, should make it possible for non-experts in natural language processing, to get started with similar practical information extraction tasks in other medical domains. This work is structured as follow. We first introduce automated information extraction in Section 1.1, after which Section 1.2 introduces the use-case of our library for prediction of *time-to-delivery* for patients at risk of Preterm Birth (PTB) at the Ghent University Hospital and presents several challenges introduced by free text from this care center in this domain. We then provide related work in Section 1.3. Section 2 collects all the used methods. Starting with details on the used dataset (Section 2.1) and challenges strongly related to the data format (Section 2.2), the proposed clinical NLP-pipeline is described (Section 2.3), with details on the individual building blocks and their functionality. After that, the used features and machine learning model for the clinical risk prediction use case are described (Section 2.4). Section 3 provides the obtained results, in particular for the experiments on bootstrapping clinical entity recognition (Section 3.1) and birth risk estimation (Section 3.2). A discussion of these results is provided in Section 4, covering an error analysis of the entity extraction task (Section 4.1), limitations of clinical information extraction (Section 4.2), and on the added value of interpretable IE features to improve clinical prediction models (Section 4.3).

1.1. Automated information extraction

In order for a decision support system to make use of the potentially rich information available in the form of natural language (e.g., in clinicians' notes or lab results), an information extraction step is needed where chunks of structured information are extracted from the unstructured data.

Information extraction is commonly recognized as a specialized area within the broader field of natural language processing (NLP) and refers to the automatic extraction of concepts, entities, and events, as well as their relations and associated attributes [5].

Because of a number of unique characteristics of data in the clinical domain, that differentiate it from the general domain or scientific biomedical literature, a focused effort is required, as will be discussed further in this work.

We adapt and apply several information extraction tools to medical notes generated during hospitalizations of pregnant patients at risk of preterm delivery. We describe and assess specific components for note de-identification, and the extraction of measurements of a clinical parameter expressed in unstructured and semi-structured medical notes. We further evaluate the effect of including the extracted features in predictive models for PTB. We show that preprocessing and extracting information from medical notes has the potential to significantly increase the effectiveness of a clinical risk prediction model while preserving model interpretability and transparency.

While our application focuses on decision support for PTB risk prediction, we hope that, by open-sourcing re-usable components, we can speed up the process of bringing powerful information extraction and machine learning models to clinical decision support models in other clinical domains. To the best of our knowledge, this is the first use case of information extraction designed for decision support models in the context of PTB risk.

1.2. Decision support for preterm birth

Preterm birth (PTB) is defined as giving birth before a gestational age of 37 weeks, as opposed to the expected gestation of 40 weeks. Globally, PTB occurs in 11 percent of all pregnancies and is one of the leading causes of death among children younger than five years according to the World Health Organization [6]. PTB can cause severe morbidities such as respiratory distress syndrome, bronchopulmonary dysplasia, necrotising enterocolitis, intraventricular haemorrhage, retinopathy of prematurity, and sepsis. Moreover, PTB can have lifelong effects on neurodevelopmental functioning such as increased risk of cerebral palsy, impaired learning and visual disorders, and is associated with increased risk of chronic disease in adulthood [7]. In Europe, the prevalence of PTB ranges from 5% in Scandinavian countries up to 11% in Austria, and the overall rates are yearly increasing in some countries. For tertiary care centers, in which our study is situated, this can be significantly higher: 18% of the deliveries in Ghent University Hospital are preterm.

Most PTBs are due to spontaneous labor and preterm prelabor rupture of membranes (PPROM) but PTB occurs for a variety of reasons. Approximately one third is iatrogenic, meaning that for medical reasons the delivery needs to occur prematurely. Common pathologies leading to iatrogenic PTB are intra-uterine growth restriction and blood pressure related complications of pregnancy. Factors associated with PTB include multiple gestations, infections and chronic conditions such as diabetes and high blood pressure; however, often no cause is identified [8]. Mortality and long-term complications can be prevented with cost-effective interventions, however, because the pathophysiology and etiology of preterm labor are not yet fully understood, deciding timely on the appropriate intervention is hard. In this setting, clinical decision

support models can be important for helping clinicians to identify women at higher risk of premature delivery, so that they can offer prophylactic interventions and help guide antenatal management decisions, much of which depend on the estimated time-to-delivery [9, 10, 11, 12].

While medication to stop contractions has little guarantee to prevent PTB, it can allow maternal administration of corticosteroids for fetal maturation as well as transfer to a tertiary centre. These measures reduce mortality, disability and intensity of neonatal care required [8]. Important for adequate treatment and prevention of preterm labor is the accurate estimation of the actual *time-to-delivery* at the moment of admission to the hospital. To assist clinicians with estimating this important variable, we leverage machine learning models to detect whether or not delivery is likely to occur within 24 hours, 48 hours or 7 days after hospitalization. These time windows are especially significant for timely administration of corticosteroids, as these are believed to have optimal effect between 2 and 7 days after administering [13, 14]. We rely on information extracted from clinical text to provide the decision support systems with additional highly informative features.

Our study was performed in the context of the 'Predictive health care using text analysis on unstructured data project', funded by imec in Flanders and the PRETURN (PREdiction Tool for prematUre laboR and Neonatal outcome) clinical trial (EC/2018/0609) of Ghent University Hospital. One of the goals of this project was to investigate the use of information extraction for clinical risk prediction and to provide guidelines for future applications.

1.3. Related work

A considerable amount of prior work presents methods to extract structured information from unstructured medical data. While some works focus on the development of general-purpose tools to create structured databases from text, others are domain-specific and intended for application. We situate our work among the latter.

Applications of clinical IE are typically related to diseases, drugs, or clinical workflow optimization. The most common application in disease studies is cancer [15], venous thromboembolism [16], peripheral arterial disease [17], and diabetes [18]. Recent applications show a trend to leverage IE to look further into refined diseases or events with specific features. Sohn and Savova [19] developed a set of logic rules to improve smoking status classification. Urbain et al. [20] mined heart disease risk factors in clinical text with named entity recognition and distributional semantic models. Topaz et al. [21] mined fall-related information in clinical notes and compare rule-based and novel word embedding-based machine learning approaches. Mantas et al. [22] detected adverse events in neurosurgery from written documents. Nasif et al. [23] built breast cancer classifiers that can help in early detection of malignancy by mining concepts from mammography records. Around the time of submission of this work, a named entity recognition module for free text in Electronic Health Records, called Med7, was released [24]. Their system achieves high accuracies on recognizing seven different

entities, and allows for transferring to other datasets with minimal fine-tuning. In contrast to our proposed work, Med7 only focuses on the entity recognition, and performs no semantic parsing or relation extraction.

Approaches to clinical IE generally involve rule-based methods, machine learning based approaches, or hybrid methods that combine both [25]. Rule-based IE systems typically consist of a set of manually engineered rules and an interpreter to apply the rules. IE approaches leveraging machine learning algorithms have recently gained interest due to their effectiveness and success in shared tasks which evaluate specific sub components of IE pipelines [26]. However, these systems generally require large annotation efforts by experts. To the best of our knowledge, we present a first hybrid IE system specifically designed for decision support models in the context of PTB. For a recent comprehensive overview on the topic of clinical IE, we refer readers to the review by Wang et al. [25]

Recently, a number of open-source libraries for NLP have emerged, with spaCy [27] being one of the most popular due to its speed, ease of use, and performance that resembles the current state-of-the-art. spaCy, out-of-the-box, includes a limited set of components for the English languages, additional packages or models, for different languages or domains, are distributed by contributors and can be installed as Python packages. Following recent work on open source packages for medical and biomedical text [28, 29], we built our tools as an extension to the spaCy library. The capability of our extension to extract semantic frames of medical events, which enable further feature extraction for use in clinical decision support systems, sets our pipeline apart from other IE systems in the medical domain and from the standard spaCy library. One of the benefits of this approach is straightforward integration with the large ecosystem of Python libraries for machine learning and other libraries based on spaCy.

Several efforts have already been made to assess the potential of predictive models of PTB risk [30, 31, 32, 33, 34]. These models are based on a large number of variables, including the gestational age, clinical history, cervical length, blood pressure, results of biomarker tests, and many others. Preferably, these are all available in a structured format, to be used as features in a prediction model. In practice however, there may be missing values in the parameters recorded in the electronic health records, or they might simply not be measured. The unstructured text fields may however contain some of the missing required variables. There are different ways to leverage unstructured text fields in prediction models. Traditional text classification methods would use all of the text, with bag-of-word features for all occurring words, meaning that only the occurrence of individual words is used, whereas word order is discarded entirely. The resulting predictions are however often not interpretable due to the unintended modeling of confounders, as will be shown in Section 3.2.2. We therefore advocate the alternative, i.e., to extract well-known and interpretable features in a structured format from the text fields, to be directly used in prediction models.

2. Materials and methods

This section presents the methods and techniques used in the underlying work. In Section 2.1 we introduce the *PRETURN* (PREdiction Tool for prematUre laboR and Neonatal outcome) dataset, containing a large collection of free-text notes related to high risk pregnancies. Moreover, we elaborate upon the task that we are trying to solve with this dataset. In Section 2.2 we highlight a number of challenges due to the style and formatting of this type of medical notes. In Section 2.3 we describe the different stages of our proposed IE pipeline which modifies and extends the standard spaCy pipeline to tailor it towards this specific style and formatting, and include functionality to extract information for inclusion in the predictive model. In Section 2.4 we discuss how the used dataset and discussed IE system are used in order to create predictive models that estimate the risk of birth.

2.1. PRETURN dataset

The original data is retrieved from a database by the Department of Gynaecology and Obstetrics at Ghent University Hospital that was constructed through the usage of an electronic health record software package. All medical notes used in this study were written and processed in Dutch. For purpose of illustrating some of the notes' characteristics, we translated these notes to English as closely as possible. Standard EHR software allows the clinicians to take notes in the form of free text or in a semi-structured format by filling in a pre-defined template. In total, 42 templates, corresponding to various clinical events and treatments, are supported. In the remainder of this paper, we refer to these templates as note types. Because different types of clinical text each have different purposes, they are highly heterogeneous in their content and level of detail. Moreover, the note types often allow free-text comments as well in order to ensure enough flexibility, resulting in a wide variety of ad-hoc constructs [35]. Because patients can have multiple pregnancies, and multiple hospital admissions per pregnancy, records are identified and linked based on patient, pregnancy, and admission identifiers.

The goal of the *PRETURN* data is to allow for the construction of predictive models for preterm birth risk estimation. In this work, risk estimation is performed by training binary classifiers which provide a probability for whether a patient will deliver within a certain amount of time. In total, three different time windows are used: (i) within 24 hours, (ii) within 48 hours, and (iii) within 7 days. These time windows were decided in consultation with experts, as these are the bounds between which the effect of corticosteroids is thought to be optimal [13, 14]. Several sources of data are available to base the estimates on. On the one hand, EHR data that is available shortly after admission, which remains static throughout the entire admission, can be used. On the other hand, temporal data which arrives in the system during the admission becomes available. This temporal data consists of structured lab results and notes taken by clinicians which are processed by our information extraction system, which is the focus of this study.

Medical notes for 3,611 patients are included in this study, corresponding to 4,332 pregnancies and a total of 5,030 admissions between 2012 and 2017. Patients at a

Data	Value
# Pregnancies	4,332
# Pregnancies (between 24 and 37 weeks of gestation)	949
# Medical notes	342,833
# Free text notes	51,082
# Semi-structured notes	291,751
# Types of notes	42
Size of Vocabulary	51,872
Average # notes/pregnancy	74
Average # token/note	26

Table 1: Overview of the preturn dataset.

gestational age less than 24 weeks are not included, since neonatal intensive care is not started before this term in Ghent University Hospital. Patients arriving at the hospital after 37 weeks of gestation are no longer at risk for PTB and do not require potential preventive measures, and are therefore not included either. After filtering, 1,065 pregnancy-related admissions are kept, corresponding to 949 pregnancies of 911 women in between 24 and 37 weeks of gestation. We summarize these properties in Table 1. From the 42 types of notes, the most frequently occurring are the semi-structured descriptions of medication administrations (21.1% of all notes), notes which are fully unstructured (i.e., 'free text' notes 15% of all notes) and notes logging vital signs and lab results.

Next to the unstructured medical notes, a number of structured EHR data are available at time of hospital admission, including but not limited to: (1) number of fetuses, (2) age (mother), (3) gravidity, (4) parity, (5) length (mother), (6) weight (mother), (7) BMI, (8) gestational age at admission, (9) duration ruptured membranes, (10) method of conception, (11) smoking history, (12) alcohol usage, (13) drug usage, (14) history of cesarean section, (15) ethnicity (mother). After encoding of categorical features, a total of 112 structured features are available for each patient. This set of features is available at the time of each admission and remains static throughout the entire admission. It is important to note that due to the fact that a patient can be admitted several times during the same pregnancy, some variables such as the gestational age or drug usage can change when a new admission occurs. The linked collection of unstructured and structured data represents the *PRETURN* dataset.

2.2. Challenges

Clinical narrative is often generated under time pressure, using a combination of ad-hoc formatting, chunked words which could be inferred from context, with heavy use of jargon and acronyms, all of which increase the information density. Based on the categorization of challenges for clinical IE systems by Leaman et al. [35], we display and illustrate some of these challenges in Table 2, some of which are common for clinical text and obstruct automatic processing as will be discussed later. Table 2 demonstrates

why IE systems cannot rely on correct grammar and writing which is common for other types of written text, such as published texts in the media sector.

2.3. Information extraction pipeline

Information extraction systems commonly involve a number of subtasks: tokenization, sentence segmentation, named entity recognition to identify concept mentions or entity names from text (e.g., person names or locations), and relation extraction to identify relations between concepts, entities, and attributes (e.g., person-affiliation and organization-location) [5, 25]. In the clinical domain, extractions aim to provide a formal representation of the clinical data [36]. In this section, we describe the different stages of our proposed IE pipeline.

A flowchart of our system can be found in Figure 1. In an initial stage, the medical notes are de-identified in collaboration with an expert (Section 2.3.1). Using these anonymized medical notes in combination with domain knowledge, the Named Entity Recognition (NER) models are bootstrapped (further detailed in Section 2.3.6), in order to generate noisy but automatically labeled data. After this phase, all the necessary inputs are generated to train our information extraction pipeline. The pipeline first tokenizes the notes, which we discuss in Section 2.3.2. Then, entities are recognized from the text, on which we elaborate in Section 2.3.3. The entity recognition model is trained using data manually labeled by an expert as well as data automatically labeled by our NER Bootstrapping component. After recognizing the entities, they are normalized to have the same representation, and linked to semantic concepts available in an ontology (Section 2.3.5). As a final step of our information extraction pipeline, relations between the recognized entities are inferred. The output of this pipeline are semantic frames, from which features can be extracted that can be concatenated with the admission information in order to provide updated risk scores and explanations to the medical team. We further discuss this decision support system in Section 2.4.

Figure 2 illustrates the individual steps of the overall system applied on an example note "No VWV or VBV, Dafalgan (1g) 1x/d". The original text (in Dutch, but translated to English for illustration purposes in Figure 2) is first split up into tokens. For example, ''1x/d" (short for "one time per day") results in a sequence of 4 tokens. After tokenization, clinical concepts are identified. For example, "Dafalgan" is identified as an instance of the semantic type *Medication*. In the next stage, clinical concepts are linked to a medical ontology by the entity linker module. Finally, in a semantic parsing step, relations between clinical concepts are identified. For example, the dose of a certain medicine is linked to the corresponding medication concept.

The following paragraphs contain a more detailed description of these pipeline components (tokenization in Section 2.3.2, clinical named entity recognition (NER) in Section 2.3.3, linking in Section 2.3.4, and finally the semantic parsing step in Section 2.3.5). However, we start with a data pre-processing step which is essential in automated processing of clinical data: de-identification of the data (Section 2.3.1). In particular, we describe our de-identification process during which data engineers had

Category	Sub-category	Example
Flexible Formatting	Variable semantics	*Section header*: "Admitting Diagnosis: SPLENOMEGALIA"
		Inseparable phrase: "Neuro: nonfocal"
	Structure w/o sentences	"Hb 11.9; L 7.13: Tc 86 (173); UZ 8.2; AST 36; ALT 31; LDH 266"
		"OXYTOCICA - oxytocine (Syntocinon) , rem: 5 IU in shot "
	Missing punctuation	"Boostrix No OGTT Maternity fee, married"
		Periods: "headache, no nausea or vomitus"
	Parenthetical expressions	"Application blood:(left arm)"
		"Boostrix vaccintation prescription present (no GP)"
Atypical Grammar	Missing expected words	*Verb*: "No BH contractions"
		Object: "Restless, probably had gastro-enteritis a while ago"
	Strange POS combinations	*Adjective without noun modified*: "Head, eyes, ears, nose, and throat examination revealed normocephalic and atraumatic."
Rich Description	Variety of textual subjects	*Patient*: "pelvis- and backache, can not sleep"
		Anatomy: "no blood loss, some fluid loss last night."
		Test or procedure: "Task nr 5.3: ketones ++"
		Family: "No relatives with diabetes"
	Variety of styles	*Diagnosis*: "Glucose measurements under control (maternal diabetes)"
		Evidence: "She had a teststrip of 142."
	Context-specific language	*Jargon*: "Uterus: 2 fingers w.r.t. umbilicus , hard"
		Ad-hoc acronyms: "sugars good, KB+, VWV- VBV-HB sporadically"
		Abbreviations: "CTG/ 2x reactive, toco flat, G2P3"
	Misspellings	"Couging [sic] since a couple of days."
		"No vaginal blooloss [sic] or stumach aches."
		"No fever. Normal mictie- and defaecationpattern. [sic]"
		"Dayly application of utrpogestan [sic] because of preterm partus."

Table 2: Illustrative examples of common challenges in processing text from clinical narratives related to PTB. For illustration purposes, notes shown in this table were translated from Dutch to English by the authors where possible.

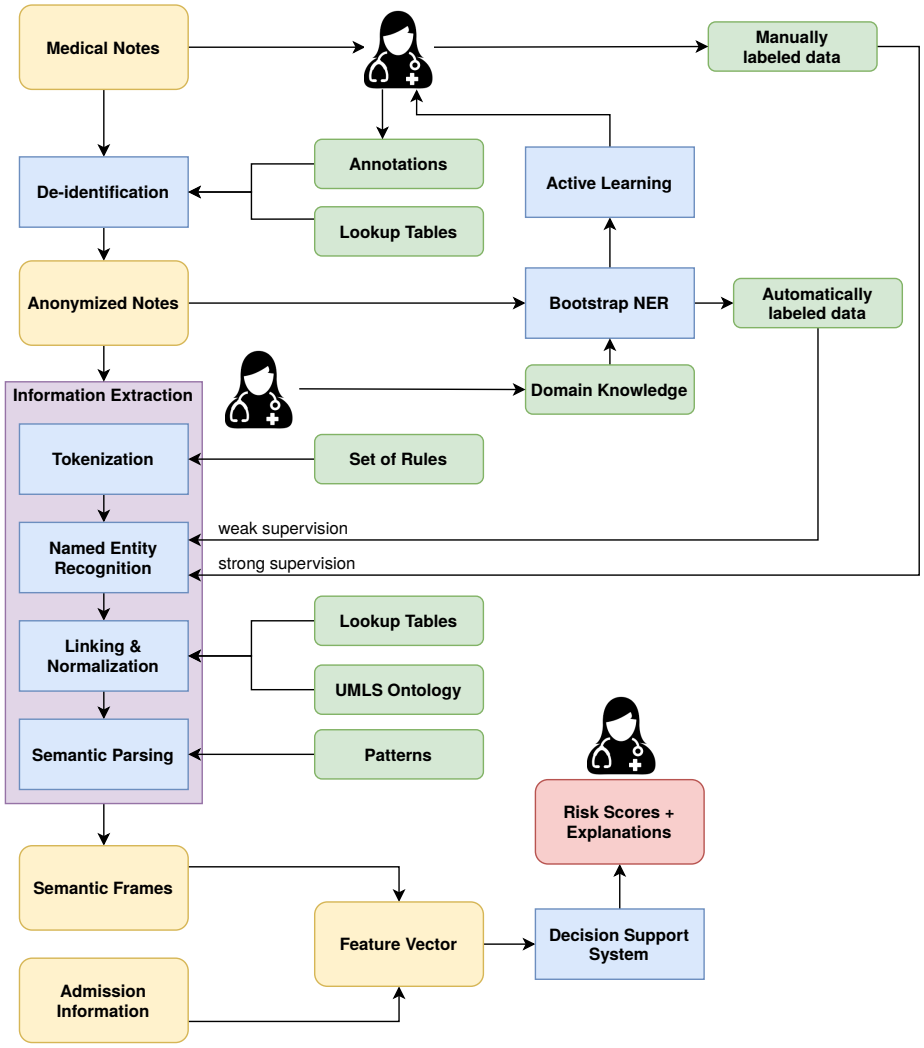


Figure 1: Flowchart and general architecture of the NLP pipeline for the presented decision support task.

no access to the original *PRETURN* dataset for developing the de-identification tools.

Implementation of our pipeline is based on the popular spaCy library for python. By default, the library includes many NLP functions to process text for a wide variety of languages, such as tokenization, part-of-speech-tagging and NER but is mostly tailored towards written media or web text. The accuracy of these components depends on the domain and amount of training data for each language. Similar to recent work on open source packages for medical and biomedical text [28, 29], we modify and extend the library for our use case. We modify two of its standard components: tokenization and NER, and include three new modules: de-identification, entity linking, and semantic parsing.

2.3.1. De-identification

Clinical notes contain detailed information about patient-clinician encounters in which patients confide not only their health complaints, but also personal choices or possibly stigmatizing conditions, all of which may be highly sensitive. This confidential relationship of medical data is legally protected in the European Union. Conditions for scientific usage of health data are set out in the General Data Protection Regulation (GDPR) [37]. The GDPR lists general principles relating to processing of personal data, including that processing must be lawful (e.g., by means of consent), fair and transparent. It must be done for explicit and legitimate purposes, and the storing of data should be kept limited to what is necessary and only as long as necessary. The GDPR excludes *anonymous* data from its scope of application under the condition that re-identification of individuals from the data is impossible. Therefore, de-identification is an important step before feeding the text to the NLP pipeline. This includes removal of any kind of protected health information (PHI) including names, locations, contact details, identification numbers, specific dates and times. Since the GDPR does not provide any strict rules about which types of PHI should be removed during de-identification, we base our PHI tagging scheme on the guidelines defined by the US HIPAA regulations. In order to de-identify the *PRETURN* dataset, the sensitive information of medical notes for 20 patients, that signed an informed consent, was highlighted by a medical expert (I.D.) and used to develop a de-identification module.

Based on these annotations, a collection of rules specifically tailored to detect sensitive information with a near-perfect recall was generated. Lookup tables, decision rules and fuzzy string matching were used to implement a rules-based de-identification step. Rules include on the one hand regular expressions to detect certain structures within the notes (e.g. phone numbers and birth times), and on the other hand binary search trees containing sensitive information such as doctor names, Belgian cities, etc. Regular expressions were created to remove identification numbers. Tables of person, personnel and location names were provided by the hospital. After achieving a near-perfect recall on the annotated set of notes, the de-identification was applied by the medical expert (I.D.) to a large collection of notes. Results were checked and errors were compiled in order to improve the rules over several iterations.

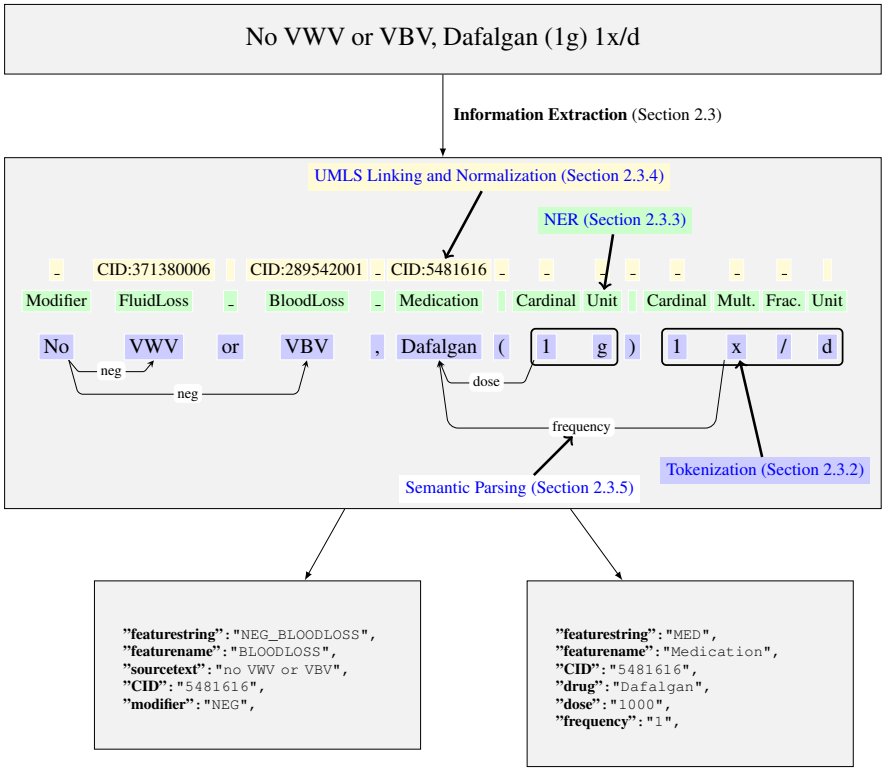


Figure 2: Example of a medical note processed at each step of the NLP pipeline.

2.3.2. Tokenization

In the first stage of our NLP pipeline we tokenize clinical notes. Tokenization is the process of demarcating sequences of characters into linguistic units called tokens, which typically correspond to words. While this operation is well supported by many NLP libraries, most are developed for prosaic text. Clinical text includes many features which are not properly handled by these standard implementations. Frequent edge cases include contractions, hyphenated words and larger constructs such as results for medical tests which include numbers, units and fractions in a single sequence of characters. Previously, we highlighted some of these cases in Table 2. While this component merely returns lists of tokens, it directly affects all later processing steps in the pipeline. During development, proper tokenization was crucial for proper information extraction. Building on a standard language-specific tokenizer, many exceptions were included to allow for the correct segmentation of informative special cases, most notably for extractions of numerical values such as "*G2P3*". We included a number of exceptions to spaCy's standard tokenizer module to guarantee proper detection of valuable numerical data. The majority of exceptions were included to separate unit measures from numerical values.

After processing all notes, structured as well as unstructured, and de-identification (described in Section 2.3.1), a vocabulary of 51,872 tokens was constructed. Our spaCy pipeline for Dutch medical text (named *nl_core_med_sm*, following spaCy naming convention) has a smaller vocabulary than that included in the related scispaCy [28] and includes pre-trained word embeddings using the word2vec library. Table 3 compares our *nl_core_med_sm* with the pipeline released by Neumann et al. [28] (*en_core_sci_sm*), trained on English biomedical text from PubMed Central Open Access [38].

Model	**en_core_sci_md** [28]	**nl_core_med_sm**
Vocabulary size	101,678	51,872
Minimum word frequency	20	2
Minimum document frequency	5	1
Processing time per note	10 ms	5 ms

Table 3: Model statistics for nl_core_med_sm and the related library scispaCy, en_core_sci_md [28] trained on biomedical text from Pubmed Central Open Access [38].

2.3.3. Named entity recognition

Clinical findings, diseases, procedures, body structures, and medications recorded in the medical notes constitute valuable information. Named Entity Recognition (NER) in clinical notes identifies mention spans (i.e., potentially ranging over multiple tokens) of the clinically-relevant concepts such as names of medications or body parts. For extraction of concepts from free text notes we include a machine learning approach to NER, whereas for semi-structured notes we can really on notes' template to extract relevant information. Machine learning methods formulate the clinical NER task as a sequence labeling problem that aims to provide the best label sequence for a given sequence of tokens.

While standardized ontologies such as UMLS and SNOMED CT [39] provide an extensive categorization of semantic types and vocabularies, these are not tailored towards use in machine learning models. Some of these semantic types are too coarse while others are too specific or sparse to be effective for using in predictive models. For optimal effectiveness in our clinical prediction model and verification by clinicians, we defined and refined (over several iterations) a custom ontology of semantic types in collaboration with medical experts, tailored specifically for PTB risk prediction. While there is no one-to-one mapping between our ontology and broad ontologies such as UMLS, most of our semantic types can be mapped to one or multiple UMLS concepts. Clinicians identified 62 different semantic types as being informative for risk prediction. These include, besides standard clinical concepts, other important concepts such as units (temperature, volume, quantity) and named entities (patient, partner, family, ...). After annotation of 3,381 documents, 21 types received over 100 annotated concepts. For a statistical approach to our NER task, this appears to be a lower bound for training a decent sequence labeling model. Therefore, for these 21 types, we use the statistical NER approach for these, whereas we use simple pattern matches based on lookup tables for the remaining 41 types. We present these semantic types, together with descriptive statistics and matching UMLS concepts in Table 4. Note that for transparency in the results for the semantic type extraction (Table 8 and Table 9), we only report test metrics for the 21 *trained* types.

For sequence labeling we rely on the statistical sequence tagging model included in the spaCy library. The NER model in spaCy is a transition-based system based on the chunking model by Lample et al. [40] Each token is represented as a hashed, embedded representation of the prefix, suffix, shape and lemmatized features of individual words. spaCy's NER model is a deep convolutional neural network with residual connections, and a transition-based approach to named entity parsing. We refer readers to work by Goldberg et al. [11] for a more detailed description of this model. While this architecture is not guaranteed to provide the best possible results for our use-case, we focus on an initial evaluation of our annotated data and schema, rather than pursuing more complex alternatives.

To increase efficiency of the annotation effort, we use models trained on the available data and phrase lists mined from UMLS to *pre*-label data. Instead of annotating documents from scratch, annotators edit the proposed annotations or add missing ones. We describe this approach in more detail in Section 2.3.6.

2.3.4. Concept normalization and linking

After the NER stage, normalization and linking of clinical entities or concepts is performed. The entity linker module links concepts, labeled with one of our PTB-specific semantic types during NER, to one of the biomedical concepts defined in the Unified Medical Language System (UMLS) Metathesaurus. UMLS distributes terminology and coding standards to enable interoperable biomedical information systems. We make use of its main component, the Metathesaurus, which organizes concepts and links similar names for the same concept for other vocabularies, as a reference[41].

For example, the acronym "MD" has 42 different meanings (Medical Doctor, Major depression, Mitral disease, etc.) according to UMLS. An entity linker decides which is the actual meaning associated with the acronym "MD" as used in the considered context, and links the considered "MD" mention to the corresponding UMLS Concept Unique Identifier (CUI).

Note that our NER model detects concepts and labels them according to our PTB-specific semantic types, after which the entity linker assigns a matching UMLS Concept Unique Identifier (CUI). UMLS, with over one million of biomedical concepts, includes a much finer grained ontology than our PTB-specific ontology. Next to a PTB-specific semantic type, also having access to the corresponding UMLS CUI, will enable use of the additional information included in UMLS such as synonyms as additional source and allow for more fine-tuning during feature engineering during development of the risk prediction models as described in Section 2.4.1. For example, the semantic type assigned to all medication types might be too granular to be useful for prediction, instead the potentially more specific UMLS CUI of the medication could be a more informative feature.

In our pipeline we include the same entity linker used in scispaCy [28]. This entity linker measures the similarity between the concepts extracted by the NER component, and compares them to those stored in UMLS using an approximate nearest neighbours search. We generate candidate UMLS concepts for each of the extracted PTB-specific concepts with a commonly used method for information retrieval, which included the following two steps. We first index all of the canonical names stored in UMLS using vector representations of sequences of three characters. Then, we employ a nearest neighbor search to retrieve the most likely candidate UMLS concept for each PTB concept. The UMLS concepts for which a concept name does not appear at least once in our corpus, are excluded from the lists of candidates. As a result, we use a total of 11,416 UMLS concepts out of a total of 287,839 CUI stored in the UMLS for Dutch. Currently we do not train this linking component and choose the most similar canonical name in the UMLS database. While this introduces noise, especially for abbreviations, we believe the effect of this noise is reduced by the highly domain-specific language related to preterm birth, i.e., apart of acronyms and single character expressions, most of the vocabulary is context independent and can be linked to a single UMLS concept. We normalize numerical words in mentions and concepts to their corresponding Arabic numerals, as well as attributes and qualifiers using manually engineered lookup tables. Furthermore, each type of unit is normalized to the same constant using lookup tables, e.g., mass units in medications are expressed as milligrams. In Figure 2 this converts '1 g' to the numerical value of 1,000.

2.3.5. Semantic parsing

Semantic parsing converts language to a machine-understandable representation of its meaning [42]. What we mean by semantic parsing in this context, is the task of identifying relations between the clinical concepts and attributes such as, e.g., medication and dosage or pain and anatomical location. Based on the semantic type assigned during the NER step, clinical entities are connected to attributes: qualifiers, temporal modifiers, measurements, and anatomic location. Relation candidate pairs are extracted

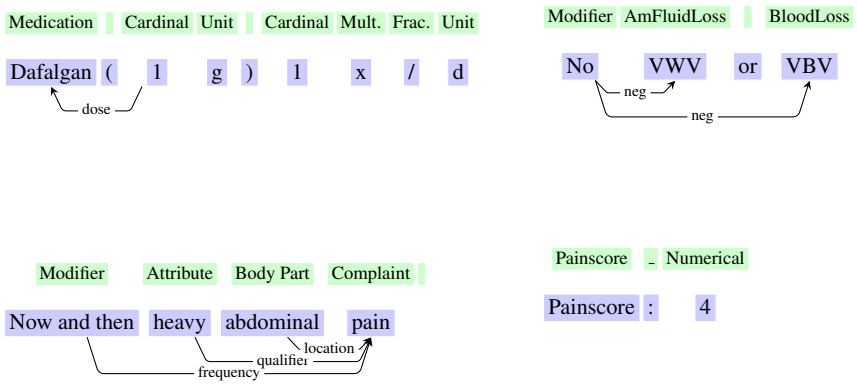
Semantic type	Example	UMLS CUI's	Train/Dev/Test	Norm. Freq
Attribute	-, normal, +, left, assisted, weak, hard, none	(Normal-C0231683), (Left-C0443246), (Hard-C0018599), ...	1019/147/127	0,382
Qualifier/Temporal Modifier	no, full, not, strong, frequent, right, standing, broken, decreased,...	(Full-C0443225), (Not-C1518422), (Spontaneous-C0205359), (Standing-C0231472), ...	985/120/157	0,373
Fraction	per, /, every,...		446/53/60	0,165
Patient	patint, pte, mama, pat, patient, ms.	(patient-C0030705), (Mama-C4209064),...	430/66/63	0,165
Time unit	week,d,day,min,hour,w,',h,u,dagen	(Day-C0439228), (Hour-C0439227), (min-C3813700), ..	424/60/69	0,164
Medication	dafalgan, perfusalgan, diclofenac, paracetamol, oxytocine, syntocinon, loramet	(Diclophenac-C0012091), (Paracetamol-C0000970), (Syntocinon-C0592155), ...	371/63/58	0,146
Medical Test/Procedure	ctg,vaginal examination,toco,glycemie,blood collection,actimpartus	(Tocography-C0040345), (Venography-C0031545), (Cardiotocography-C0007208), (Glucose strip test-C4761113) ...	379/68/44	0,145
Cervix	cervix,cx, cxv	(CERVIX-C0007874)	340/63/68	0,139
Range	-, to,ranging from, between		288/55/32	0,111
Complaint/Pain	pain, tired, complaints, hemorrods, stress	(Pain-C0518090), (Tired-C0015672), (complaints-C0277786),	290/38/44	0,110
Blood Loss	vag bvl, bloodloss, vbv, blood loss	(Blood loss-C0019080), (Vaginal Hemorrhage -C2979982)	202/27/35	0,078
Amniotic Fluid Loss	vwv, amniotic fluid loss, vag vvl	(Amniotic Fluid Loss-C0238625),	155/21/23	0,059
Length Unit	m, cm, mm, fingers..		133/16/19	0,050
Mass Unit	kg, g, lepels, ...		118/30/17	0,049
Baby	baby, child, son,	(Baby-C0021270), (Child-C0008059),	111/17/28	0,046
Sleep	resting, sleep, nap	(Sleep - C0037313), (Rest - C0035253)	107/23/18	0,044
Contractions	contractions, cramps, hb	(Cramps-C0026821), (Uterine Contraction-C0042130)	105/18/21	0,043
Volume unit	ml, l, centiliter cc, ...		106/25/8	0,041
Way of Application	oral, intraveneus, rectal, iv, vaginal, intramuscular	(Rectal-C0205052), (IV-C0022326), (Vaginal-C0042232), ...	92/15/16	0,036
Condition	HIV, diabetes gravidarum, zws diabetes	(HIV-C0019682), (Diabetes-C0011847)	89/10/21	0,035
Cardinal Numbers	seventy-nine, fifteen, twenty		92/13/8	0,033

Table 4: Overview of semantic types in the NER module and number of annotations in training, development and test collections. All shown examples of annotated concepts were translated from Dutch to English by the authors where possible.

by comparing clinical entities (Drugs, Complaint, Medical Test, ...) to matching attribute mentions (frequency, dose, body part, temporal constraints, ...). We apply *shallow* semantic parsing in which we the constituent parts are first identified and then linked together based on manually engineered grammar rules. Rules were developed by analyzing common sentence structures. While this approach fails to parse complex sentences, we obtain reasonable coverage, because of the limited length of notes and the prevalent semi-structured format.

Concepts of the type Medication can be linked to the concepts of route, dose, frequency and/or form. Complaints are linked to temporal modifiers and the concept of intensity. Medical tests and procedures are linked to measurements. Rules were created using spaCy's pattern matcher and stored as JSON format. Figure 3 displays several examples of patterns and parsed notes.
One challenge faced in clinical NLP in general, is that the meaning of clinical entities is heavily affected by modifiers such as negation. We rely on the semantic parsing stage to negate entities. Thus, negation detection is an important task of the semantic parsing module: to determine whether a criterion is used for inclusion or exclusion purposes. In our pipeline we use semantic parsing as means for rule-based negation detection, similar to the prominent algorithm NegEx [43].

The semantic parsing module returns frames containing all the linked concepts, identifiers and typed relations formatted as JSON records. This is the final module of the IE pipeline and *semantic frames* are generated as final output of the pipeline. An overview




```
{"label":"Diastolic","pattern": [ {"ENT_TYPE": "Pressure_Unit"},{"TEXT":"/"}, {"LIKE_NUM": true}]}
{"label":"Dilation","pattern": [{"ENT_TYPE": "MedicalTest", "OP":"+"}, {"LIKE_NUM": true}, {"ENT_TYPE": "LengthUnit"}]}
{"label":"Complaint","pattern": [{"ENT_TYPE": {"IN":["Complaint"]}, "OP":"+"}, {"ENT_TYPE": "Body_Part", "OP":"+"}]}
```


Figure 3: Semantic frames and examples of note types formatted as JSON. Green indicates the assigned semantic type.

of all data collections is shown Table 5.

2.3.6. Bootstrapping for clinical NER

As recognition of clinical entities is a key building block on which the linking and semantic parsing components rely, we conducted an intrinsic evaluation of our model. In particular, we applied and compared different strategies to generate training data and supervise the NER model (see Section 3.1). Manual annotation is a time-consuming and costly task, and requires input from clinical experts. Clever ways to increase efficiency of the *supervision* process, and thus reduce the amount of labeled data needed to train NER models, are especially useful in this setting. This section introduces the different methods that we investigated for bootstrapping the NER module without the need for extensive manual annotations.

Bootstrapping our NER models refers to training the models from scratch but leveraging existing knowledge sources and choosing a specific sample set to be labeled, in order to quickly improve the model. For example, instead of randomly choosing notes to be labeled, efficiency can be gained by presenting the most informative medical notes to annotators. A well-known set of techniques in that direction fall under the term 'active learning', whereby intermediately trained models provide clues on which samples are the most informative to be annotated. Compared to standard supervised machine learning approaches, active learning offers the opportunity to build classifiers with a reduced amount of manual annotations, especially for learning to recognize rare types of concepts [44].

Next to training on fully labeled notes, which we refer to as *strong* supervision, we include sources of *weak* supervision. In weak supervision, unlabeled data is annotated using phrase lists and regular expressions, leading to noisy and incomplete annotations. In our weak supervision setup, phrase lists are constructed for each of the se-

Named Entity Recognition	
# Training documents	2,581
# Development documents	400
# Test documents	400
# Tokens	69,800
# Semantic types	62
# Annotated seed terms	1,837
# Annotated concepts	13,105
Av. # Tokens/Document	20
Av. # Concepts/Document	3.9
Normalization and Linking	
# UMLS concepts	4,225,752
# UMLS concept names	14,608,810
# Dutch UMLS concept names	287,839
# Candidate UMLS Concepts names	11,416
Semantic parsing	
# Frame types	24
# Grammar rules	79
Feature extraction	
# Structured features	112
# Extracted features	10,099
Av. # Features/Document	1.3

Table 5: Descriptive statistics of data sources, annotated medical notes and engineered rules.

mantic types by mapping types to UMLS concepts and extracting synonyms for these concepts stored in the UMLS Metathesaurus. Terms matching one of the phrase lists are then automatically annotated with the corresponding type, regardless of context. This results in a significant extra amount of potentially noisy training data. This can boost performance when compared to training on the original smaller set of data having strong or gold-standard supervision. We hypothesize that the added value from weak supervision in this context comes from the relatively high quality of labeled data, because many of the annotated terms are context-independent or unambiguous, such as drug names or body parts.

Next to phrase lists mined from the UMLS database, we use word embeddings to expand a set of *seed* terms, similar to [44]. Seed terms are annotated concepts or phrases which are used to generate new lists of phrases having the same type. Word embeddings project tokens to a low-dimensional shared vector space (e.g., typically only a few 100 dimensions, compared to tens of thousands of different tokens). Neighboring words in this vector space tend to be grammatically and/or semantically similar. For each of the seed terms, the 20 nearest tokens in embedding space (in terms of cosine similarity) are manually assessed. Those confidently judged to describe the same semantic type as the seed term, are added to the list of valid expressions for that type. Presenting semantically similar tokens to annotators for evaluation is a highly efficient

way of extending term lists for semantic types. Table 6 shows 4 seed terms and the top 8 words words with the most similar representation in the embedding space (again literally translated to English for convenience).

bloodloss	**dafalgan**	**increased**	**ctg**
brownloss	daf	incremented	monitor
loss	dafal	lowered	CTg
blvl	tablet	boost	monitor
bloloss	dalalgan	reduced	CTG
slimeloss	pctm	driven up	registration
fluidloss	dafalghan	increaset	ECG
blood	Dafalgantablet	drive up	EKG
blodloss	paracetamol	recoverd	stan

Table 6: Term expansion for four seed terms word embeddings.

In total, 3,381 free-text documents from the *PRETURN* dataset are annotated according to the data schema presented in Section 2.3.3, resulting in 13,105 annotated concepts. During the process, the NER model is updated every 100 annotated documents, in an active learning setting. Each time, the next batch of most informative notes is determined by sampling based on uncertainty of the assigned labels. When a minimum of annotated concepts for a semantic type are annotated, terms are expanded using the previously described word embedding approach. Depending on note length and complexity, on average 100 notes are annotated in around 60 minutes. Presenting notes to be annotated with label suggestions predicted by the partly trained NER model allows reducing the average annotation time per note by half. After 100 notes are annotated, the NER model is trained on all available annotations and remaining, unlabeled documents are labeled, ranked and presented to the annotator for a new batch of annotations, in the next cycle of the active learning process.

2.4. Risk prediction for preterm birth

In this section, we discuss which types of features are extracted from the *PRETURN* data, and which techniques are used in order to create models that estimate the risk of birth within specified time windows.

2.4.1. Feature Extraction

The medical notes are processed by the IE pipeline in order to generate semantic frames. From these frames, structured features are extracted which are then combined with both features that were available at the point of admission and lab results that are available at that point in time. The features available at admission are described in more detail in Section 2.1. The information included in semantic frames allows for a large number of different features to be extracted. We can define three types of features:

Numerical Features include quantitative information, expressed in laboratory test results, length attributes, drug dosages, and visual analog scale scores. Visual

Description	Example	Feature
Numerical Features		
Cervical Length	Cervix length has decreased to **10mm**	10
Dilation	Dilation has increased to **4cm**	40
Gravidity	Third pregnancy of patient, G**3**P1	3
Gestational age	Note: term of **33w7d**	33.22
Blood pressure	RR: 90/120	90, 120
Heart rate	Controle ante partum: 120 bpm	120
Vomit VAS	Vomit: 3 VAS	3
Categorical Features		
Pain	Pain has increased since this morning	pain:increase
Blood Loss	No blood loss	bloodloss:none
Fluid Loss	VWV++	fluid_loss:increase
Blood Pressure Level	Blood pressure has decreased	bloodpressure:decrease
Event Features		
Rupture of Membranes	PPROM at 25 weeks	event_pprom
No complaints by patients	Patient has no complaints.	event_nocomplaints

Table 7: Examples of extracted features.

analog scale (VAS) recordings are a subjective measure mostly used for acute and chronic pain but also indicative for the level of nausea and tiredness in our study. Scores range between 1 and 10, corresponding to "no pain or "no nausea" and "worst pain or "extremely nauseous" respectively.

Categorical Features encode attributes together with clinical entities such as anatomical site, qualifications, and temporal modifications.

Event Features are binary features which indicate the occurrence of events or other conditions. They encode significant events during hospitalization such as rupturing of membranes, or patients being diagnosed with a condition.

Table 7 shows a number of extracted features of the three categories with corresponding expressions in written text. To reduce the effect of erroneous extractions, we include rules to filter out information which is highly unlikely or physically impossible. Examples of such rules are filtering out values that exceed the biological range, or information found in note types in which they are unlikely to be found.

We use labeled data from past hospitalizations to train and test predictive models. Binary and categorical features are aggregated and concatenated with the admission information for each patient at the point of prediction. Every 24 hours after hospitalization, delivery within the three time periods is predicted from that point in time by the models. For numerical features only the latest registration of the measurement is included.

2.4.2. Machine learning methods and interpretability

Feature vectors are processed by CatBoost, an implementation of gradient boosted decision trees [45]. There are several reasons that motivate this type of model and this specific implementation. First, gradient boosting algorithms are able to achieve state-of-the-art performances on structured, tabular data [46], especially when the data is

high-dimensional and many interactions between features exist. Second, CatBoost is able to automatically deal with the encoding of categorical variables and the imputation of missing values, which are both present in our dataset. Third, the CatBoost package specifically sets many of the hyper-parameters heuristically based on inferred properties of the provided data, as opposed to having static default values which is the case for other gradient boosting packages. This property, in combination with the fact that CatBoost is a very recent implementation with many of the newly found insights from recent machine learning research incorporated, results in predictive performances that exceed those of tuned models from competitive packages such as LightGBM [47], XGBoost [46] or H2O [48] without any hyper-parameter tuning [49].

One other important aspect that needs to be considered is that, especially in a clinical setting, predictive models need not only to be well-performing, but also need to be trustworthy, transparent, interpretable and explainable [50]. Clinicians need to be able to understand how any proposed algorithms may contribute to improving patient care within an interpretable workflow. Shapley Additive Explanation (SHAP) is a technique that is able provide an explanation accompanying a prediction by calculating how much each of the variables contribute towards each predictions [51]. The technique originally is model-agnostic but has a computational run-time that is exponential as a function of the number of features. Fortunately, an implementation specifically for tree-based models exists that can exactly calculate the contribution of each feature with a run-time that is quadratic in terms of the maximum depth of the tree in the ensemble [52]. Moreover, trees constructed through gradient boosting are often shallow. This results in gradient boosting trees being an excellent trade-off between model interpretability and predictive performance.

3. Results

In this section we present two evaluations of our approach. First, we provide an intrinsic evaluation (i.e., evaluating the IE quality of individual building blocks in the pipeline). In particular, Section 2.3.6 focuses on the NER stage with standard classification metrics: the precision, recall, and F_1-score, using different strategies for supervision. Next, Section 3.2 provides an extrinsic evaluation: we measure the impact of using the combined features extracted from text, ranging from clinical concepts to the semantic frames, for the clinical risk prediction model.

3.1. Bootstrapping clinical NER models

For evaluation of our clinical NER model, we split annotated free text notes into 2,581 training notes, 400 notes for development and hyperparameter tuning, and 400 notes for testing. We perform a quantitative analysis on the 400 test notes not used for training or model development. Table 8 shows macro- and micro-average precision, recall, and F_1 scores of the NER model on the test data, evaluating for exact matches as opposed to including partial overlap between ground truth and predicted mentions as correct predictions. Macro averaging calculates the mean of all individual scores for each semantic type, micro averaging aggregates all predictions and calculates one single metric. Micro averaging inherently assigns more importance to prevalent types.

Our weakly supervised approach, using automatically labeled data and heuristics, obtains significantly higher scores for recall than models trained using only strong supervision. We combine weak and strong supervision by training the NER model using weak supervision for 10 iterations and then fine-tuning the model using strong supervision. Combining both strong and weak supervision achieves a better trade-of with precision yet still obtains lower scores for F_1. Table 9 shows precision, recall and F_1 results for different models for the 21 different semantic types, on test data.

	Macro-Prec.	**Macro-Rec.**	**Macro-F_1**	**Micro-Prec.**	**Micro-Rec.**	**Micro-F_1**
Strong Supervision	49.42	52.77	51.04	47.72	50.84	49.23
Weak Supervision	43.00	77.99	**55.43**	40.01	80.27	**53.40**
Strong + Weak Supervision	51.32	52.89	52.09	49.79	52.53	51.13

Table 8: Evaluation scores for different training schemes for NER model averaged over all semantic types.

3.2. Birth risk estimation

In this section, we assess the added value, in terms of predictive performance and model interpretability, that the IE pipeline brings to the predictive models that estimate the birth risk.

3.2.1. Predictive performance

The main goal of this experiment is to assess the added value of the information extraction module in terms of the predictive performance of the model. To do this, we evaluate a CatBoost model for each of the three time windows discussed in Section 2.1 using different fixed points of elapsed time since admission. For each of these points of time, patients that already delivered are excluded and a new model is trained. The reason for re-training a new model is that the data at a certain point of time after admission are different from the data at an earlier point. In total, six points corresponding to one to six days after admission respectively are evaluated. Although the number of days between admission and giving birth in our dataset ranges from 0 to over 100 days, six evaluation points are chosen as a trade-off to show the increasing trend in performance when patients are admitted longer in the hospital while not cluttering the results. In total, this results in 18 different measurements per configuration, as depicted in Figure 4. The choice of evaluation points is in no way a fundamental limitation, only a practical choice related to clarity in reporting for this paper.

We evaluate the impact of our IE system by comparing the results of four different feature sets:

- A feature set using only admission information and numerical, structured lab results available at a certain point in time (*Baseline*).
- A feature set that includes the features from *Baseline*, in addition with bag-of-words features extracted from the free text notes (*BOW*). The latter features are

Supervision→	Strong			Weak			Strong+Weak		
Semantic Type↓	**Prec.**	**Rec.**	**F_1**	**Prec.**	**Rec.**	**F_1**	**Prec.**	**Rec.**	**F_1**
Attribute	29.19	31.97	30.52	19.50	85.03	31.73	33.54	36.73	**35.06**
Qualifier/Temporal Modifier	46.21	50.83	48.41	23.89	80.83	36.88	45.03	56.67	**50.18**
Fraction	27.96	49.06	**35.62**	20.08	98.11	33.33	29.41	37.74	33.06
Patient	65.82	78.79	71.72	57.27	95.45	71.59	67.09	80.30	**73.10**
Time unit	61.02	60.00	60.50	48.76	98.33	**65.19**	59.70	66.67	62.99
Medication	56.52	61.90	59.09	53.92	87.30	**66.67**	64.52	63.49	64.00
Medical Test/Procedure	41.79	41.18	41.48	26.75	89.71	41.22	48.33	42.65	**45.31**
Cervix	73.58	61.90	67.24	79.07	53.97	64.15	74.07	63.49	**68.38**
Range	38.24	23.64	29.21	66.67	14.55	23.88	42.86	32.73	**37.11**
Complaint/Pain	37.70	60.53	**46.46**	20.75	86.84	33.50	34.48	52.63	41.67
Blood Loss	43.33	48.15	45.61	31.94	85.19	46.46	45.71	59.26	**51.61**
Amniotic Fluid Loss	50.00	52.38	51.16	50.00	71.43	**58.82**	54.55	57.14	55.81
Length Unit	41.67	62.50	50.00	19.74	93.75	32.61	33.33	56.25	**41.86**
Mass Unit	43.33	43.33	43.33	36.99	90.00	**52.43**	36.00	30.00	32.73
Baby	50.00	58.82	54.05	42.50	100.00	**59.65**	55.00	64.71	59.46
Sleep	94.74	78.26	85.71	95.00	82.61	**88.37**	94.44	73.91	82.93
Contractions	64.71	61.11	62.86	27.27	83.33	41.10	64.71	61.11	**62.86**
Volume unit	55.56	60.00	57.69	57.14	96.00	**71.64**	62.50	60.00	61.22
Way of Application	50.00	60.00	54.55	40.00	80.00	53.33	55.56	66.67	**60.61**
Condition	12.50	10.00	11.11	35.71	50.00	**41.67**	14.29	10.00	11.76
Cardinal Numbers	53.85	53.85	**53.85**	50.00	15.38	23.53	62.50	38.46	47.62

Table 9: Evaluation metrics for different semantic types and training strategoies of the NER task on held out test notes.

binary features indicating whether or not a particular token occurs in the considered note.

- A feature set that includes the features from *Baseline*, complemented with the features extracted by the IE system (*IE*).
- A feature set that includes the features from *Baseline*, *BOW* and *IE* (*All*)

For each feature set, we carry out 5-fold cross-validation and predict time-to-delivery at the start of each day for the set of patients in the held-out test collection. Because of the limited number of patients compared to the number of features, in order to prevent overfitting, we limit the length of feature vectors to include the 100 most frequently extracted ones. We report the mean F_1 score and corresponding standard deviation as a function of the number of days since admission of our patient population. Moreover, when both the precision and recall are higher than the *Baseline*, a '+' is used as a marker. In these cases, an improvement with respect to the baseline is obtained for any clinical trade-off between precision and recall. The results are depicted in Figure 5.

3.2.2. Model interpretability

Using extracted information, clinicians can visualize what data the model "looked at" for each individual patient, which can be used to determine whether a prediction is based on credible facts, and potentially help to decide on actions. To demonstrate such potential insights gained through the automatic extraction of information, we show extracted numerical features over a period of time. One of which is the cervical length,

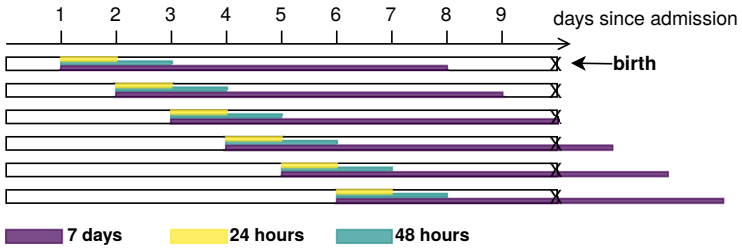


Figure 4: We evaluate our CatBoost model for three different time windows at six different points in time. These points in time are relative with respect to the admission day. In this example, the target for the time window of birth within 7 days becomes positive from 3 days since admission onwards. Targets for the first 2 days since admission are negative (i.e., birth occurs outside of the 7-day window starting on the considered day after admission).

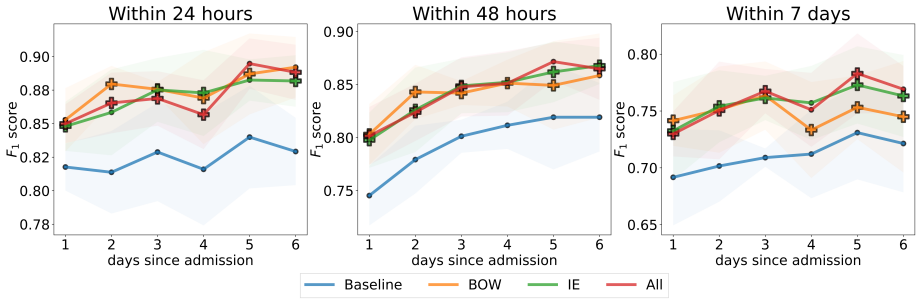


Figure 5: The average F_1 for the three time windows at the different points in time. The standard deviation is indicated by light shading around the curve.

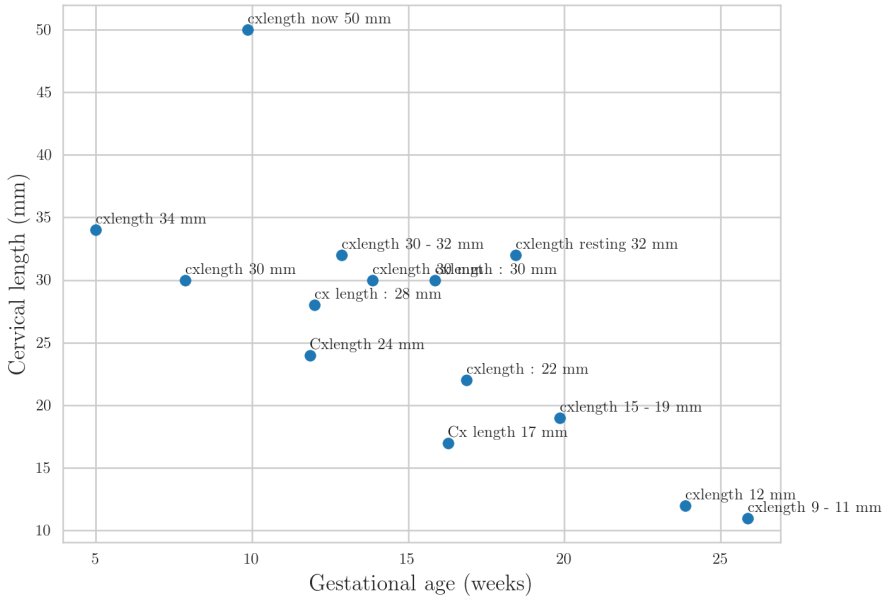


Figure 6: Extracted values for cervical length for several pregnancies; the corresponding free text is annotated next to each marker.

a crucial indicator of risk of preterm term birth. As cervical length decreases, the risk of spontaneous PTB increases [53, 54]. Results of ultrasound measurements of cervical length are often included in textual form in medical notes. Figure 6 visualizes the general decrease of cervical length over time after admittance, for a sample of several admissions at Ghent University Hospital. A complete timeline of numerical values extracted from free text notes for a specific patient using the IE pipeline is shown in Figure 7. Numerical information is shown versus the number of hours after admission.

Moreover, as mentioned in Section 2.4.2, gradient boosting allows for quick generation of an explanation corresponding to a prediction, or to quickly calculate the importance of the variables used. Table 10 shows features ranked according to the influence on the decision of the predictive model. In Figure 8, we visualize the twenty most important features using Shapley values. Each dot corresponds to one feature value of one patient. A red-colored dot represents a large value for that patient's feature, while blue indicates a small value. Gray dots represent missing values. The position on the x-axis depicts the impact on the model's prediction: the right-most dots have a positive impact on chance of birth (i.e. increase the probability) and the left-most dots a negative impact. Bag-of-words features are indicated with 'text'.

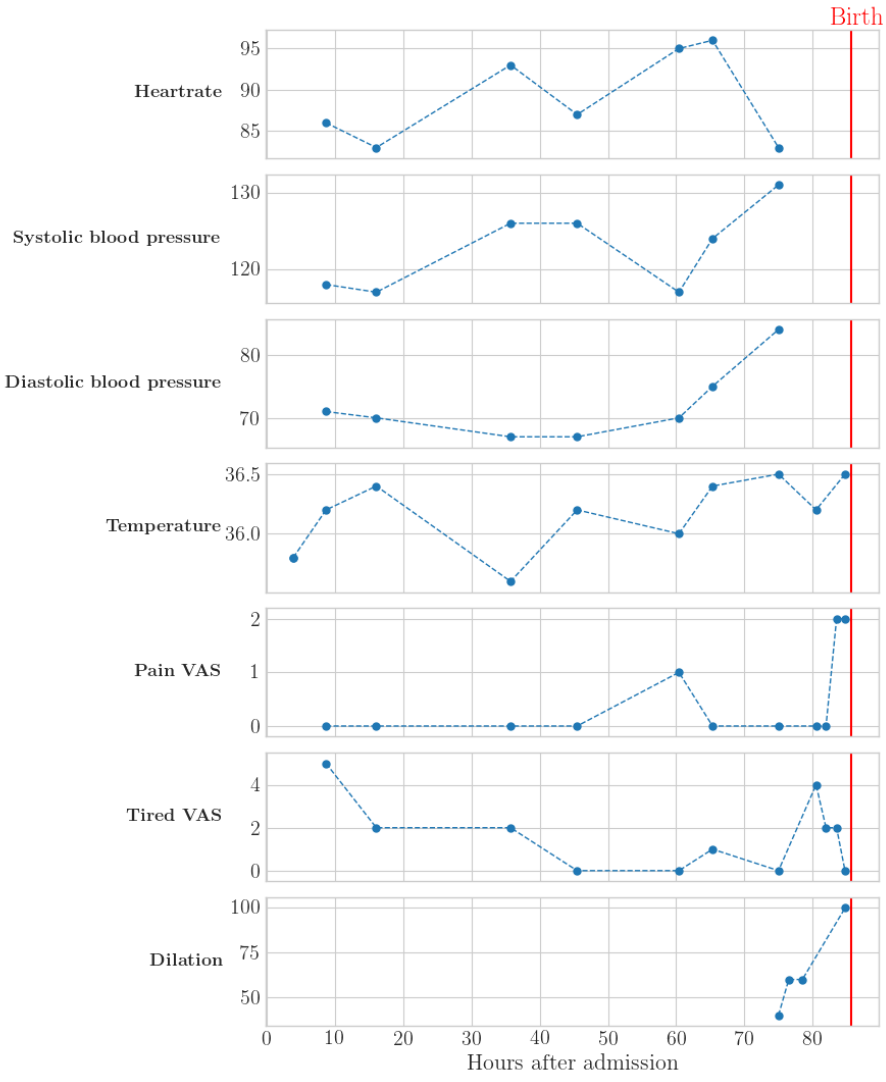
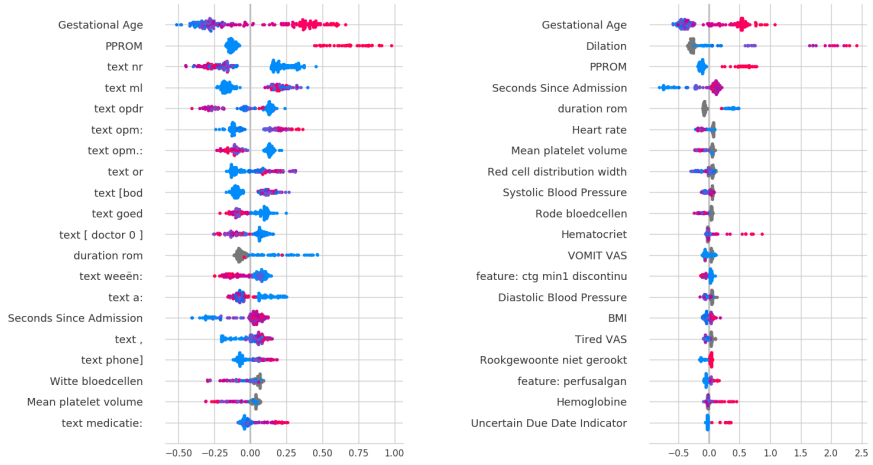


Figure 7: Extracted values from notes versus hours after admission.

Structured		Bag-of-Words		IE Features	
#	Feature	#	Feature	#	Feature
1	PPROM	2	text_membrane: broken	3	numeric:dilation
2	Gestational Age	3	text_cse	5	numeric:blood_pressure
3	BMI	3	text_weak	4	event:contractions
4	IVF	6	text_hartmann	10	numeric:tired_vas
5	Duration ROM	7	text_broken_.	14	numeric:pain_vas
6	Patient height	8	text_test:	15	numeric:temperature
7	Age mother	9	text_. amniotic_fluid:	16	numeric:heartbeat
8	Current gravidity	10	text_broken	18	categorical:cervix_unripe
9	Mean Corpuscular Volume	11	text_. weak	19	categorical:negated_complaints
10	Hemoglobine	12	text_good	20	categorical:amnioticfluidloss_decrease

Table 10: Most informative features in CatBoost models based on structured features (left), bag-of-words features extracted from notes indicated with 'text_' (middle) and features extracted using the IE pipeline (right). # = rank when sorted according to feature importance



(a) Structured + bag-of-words features

(b) Structured + IE features

Figure 8: The Shapley values for the most important predictors of the models.

4. Discussion

In this section, we perform a critical analysis of the evaluation reported in the previous section and highlight some potentially interesting future steps for our clinical information extraction pipeline.

4.1. Error analysis of clinical NER

To better identify current shortcomings of our approach for NER, we manually investigate 100 false positive predictions on the test instances by the NER model shown in Section 9. We mainly notice a prevalence of our model to over-predict the attribute and qualifier entities, as 56 of the investigated error cases involve one of these types. Especially short, symbolic attributes and modifiers such as lab results are among the most difficult concepts to extract. This appeared to be mainly because of use of single characters such as '+' or '-' to indicate positive or negative results, or the increase vs. decrease of lab measurements. Because these are difficult predictions, in our active learning strategy many of these predictions are selected in order improve the model on how to distinguish these, yet their label can remain subtle. Because of the sensitivity to these, we recommend a pattern based approach for single characters. The effect of these wrong predictions on downstream use remains limited due to the frame based semantic parsing in which these types are often used in combination. Abbreviated mentions account for another substantial proportion of the failure cases. A separate, pattern based approach which expands acronyms during a pre-processing step could aid for these cases. Other false positive predictions are harder to trace, and we recommend exploring alternative model architectures for NER, besides the standard spaCy NER model.

When considering the absolute F_1 values in Table 8 and Table 9, one could question whether they are sufficiently high for practical use. One must take into account that these represent accuracy for extractions from the more challenging free text as opposed to semi-structured text. More importantly, the goal of the IE pipeline was to extract interpretable features that contribute to better clinical risk predictions, preferably with a minimal effort (in terms of manual annotations) to develop the extraction modules. The extracted features are interpretable by construction (unlike the BOW features, as discussed in Section 4.3), and contribute to improved clinical predictions (as seen from Fig. 5). This is already a useful insight: even though the IE output may be noisy, the automatically extracted concepts do contribute to better clinical predictions. We hypothesize that during training, the clinical prediction model learns to leverage high quality features, and learns to ignore the rare, difficult, or more noisy IE features. However, due to the active learning process described in Section 2.3.6, the set of labeled notes gets biased towards notes with the most difficult cases. The test set is sampled from the set of labeled notes, as described in Section 3.1. Since the IE task was not the goal by itself, no additional test set with random samples of notes was created. The metrics reported in the aforementioned tables reflect therefore a worst case scenario, focusing on the entity extraction performance for the most difficult notes. The presented metrics are useful for relative comparisons between methods or semantic types, but the absolute levels have to be interpreted with the above in mind.

4.2. Limitations

We address a number of limitations to our information extraction pipeline and provide possible directions for future research. While our lookup-based de-identification method provides a "safety net"" and obtained a near-perfect recall after several iterations, it ignores context and de-identifies potentially valuable information. PHI instances which are abbreviations or acronyms for names are especially hard to detect and can be ambiguous as they are easily confused with medical terms and measurements. Also the generalizability of our de-identification methods across languages and domains is largely unexplored. Moreover, as all of the free text notes used within this study are in Dutch and situated within the domain of preterm birth, the generalizability of our de-identification methods across languages and domains is unexplored and uncertain. Machine learning based methods could provide better generalizability but require labeled data. Trieneset al. [55] show that a popular neural architecture generalizes best even when limited amounts of training data are available.

Our IE pipeline was evaluated on clinical notes stemming from one specific domain related to early birth risk estimation. It needs to be tested further to understand our systems generalizability. While the presented weak supervision strategies reduce the effort needed for annotating, a considerable amount of manual engineering is still required. Also, although different use cases share similar syntactic structures which help with parsing, the conditions in different fields have their own distinct characteristics. For example, the concept of "amniotic fluid loss" is common in obstetrics and gynaecology but is rarely used in diseases such as cancers, which need dedicated concepts to deal with pathology and radiology reports. We hope that, by building on an open source framework for NLP and by sharing annotation guidelines, more annotated corpora become available in the future to enable continued improvement.

During the initial stages of our research, a collection of semantic types relevant for risk prediction was constructed, in collaboration with clinicians. An alternative approach would be to extract all UMLS concepts detected by a linker or simple string matching on the collection of notes. Such bottom-up clustering of extracted UMLS concepts would lead to an alternative ontology. Benefits of this approach would include the large set of UMLS concepts to bootstrap NER models with and would allow existing UMLS concept extractors to be used (such as MetaMap[56] or CTakes[57]). We do not exclude that such alternative approach may lead to a similar gain in predictive performance. Yet we expect that mapping and validating all relevant concepts from the UMLS ontology (which contains 5 million concepts) would require many iterations of validation by experts to reduce noise.

Our approach of shallow semantic parsing allows chunking a large portion of medical notes due to their ad-hoc structure but is not suited to parse longer and more intricate expressions. A valuable extension to our pipeline would be to include machine learning based approaches for the task of semantic parsing or to implement measures that allow semantically annotated grammar rules to be generated semi-automatically.

The tasks within our proposed pipeline, such as NER and semantic parsing, are strongly dependent on each other. In this study, however, only the NER and Decision Support components of the pipeline were evaluated. There are two reasons for that. Firstly, the primary goal of this study was to leverage an IE system in order to increase the

predictive performance of the decision support system. Second, labeled data was unavailable for most IE components, with the exception of the NER task. This makes a qualitative evaluation difficult for these other components. Therefore, an important line of future work would be to leverage annotated data for different sub-tasks to train better task-specific models. In particular, multi-task learning (MTL) leverages overlapping representations across sub-tasks and is one of the most effective solutions for knowledge transfer across tasks.

While our study includes an extensive benchmark of the predictive model on historic data, it is performed retrospectively. Before application in clinical practice, a robust, prospective clinical evaluation is needed. We do believe explainable AI approaches are more likely to facilitate faster adoption into the clinical healthcare setting. As illustrated in the figures and tables in the result section, information extraction leads to more interpretable models. We hypothesize that it will help foster vital transparency and trust with the users of resulting clinical decision support tools, in high-stakes settings such as early birth risk prediction.

4.3. Birth risk estimation

By inspecting Figure 5, we observe that pre-processing notes using our information extraction pipeline results in higher scores for all time windows then models trained solely on structured features. Moreover, a bag-of-words representation of the free text notes results in an increased performance, competitive to the performance obtained with the IE feature set. Combining both feature sets does not always result in the best performance, which could be explained by a too large number of (noisy) features. The increases in F_1 score for all three feature sets, compared to the *Baseline* feature set are all statistically significant ($p < 0.05$) according to a one-sided bootstrap hypothesis test, except for the *BOW* feature set evaluated 5 days after admission on the task of predicting birth within 48 hours and evaluated after 4, 5 or 6 days of admission on the task of predicting birth within 7 days. It should further be noted that no feature selection was applied to generate these results, although this may result in a slightly further increased performance.

One other thing that can be noticed from Figure 5 is that both the feature set with a bag-of-words representation and the feature set with features extracted by our IE component have similar predictive performances. Nevertheless, using features extracted by our IE component has a significant advantage in a clinical setting. Naively tokenizing words in medical notes into sparse bag-of-words feature vectors quickly leads to uninterpretable predictions and modeling of confounding variables. This is demonstrated by investigating important features according to models trained on bag-of-words representations in Figure 8a. While some terms hint towards significant medical events such as *broken*, most are hard to interpret without additional context. One example of such a feature that has a large contribution towards the output of the model, but has minimal to no clinical interest is the occurrence of 'opm:' (Dutch abbreviation of 'opmerking' which means 'comment') having a positive impact on the final probability. In contrast, the occurrence of 'opm.:' has a negative impact. Clearly, these type of insights are of no clinical interest, and could have significant negative effects when such a model would be deployed. For example, they could be related to specific notes types, or habits

from particular clinicians in making notes, but would without doubt not be transferable to other settings (e.g., other hospitals), unlike some of the more interpretable features.

Structured features such as premature rupture of membranes (PPROM) and gestational age have a strong impact on the model's positive prediction. Out of features extracted from text, numerical features rank among the most important extracted features for the model. Prominent features for the predictive model are the cervical dilation (numerical), high blood pressures (numerical), premature rupture of membranes (event), no complaints (event) and increased heart rate (numerical). Because of their clear benefit for predictive models, our results advocate that these type of variables should be logged in a structured format. Therefore, while the IE feature set does not always consistently outperform the BOW feature set, the model is clinically more useful. Still, it is important to note that while these features are informative for model predictions they are not necessarily causal.

5. Conclusion

In this paper we studied the application of clinical information extraction to support predictive models for clinical decision support in the domain of PTB. We demonstrated that pre-processing and extracting knowledge from medical notes significantly increases the accuracy of decision support models for estimating the *time-to-delivery* while preserving model interpretability and transparency. By releasing our code[1] and documenting our workflow we hope to boost research on domain-specific information extraction to support clinical decision models. We emphasize that further research and external validation is needed to study the applicability, the cognitive impact, and the clinical utility of the presented research.

Conflict of interest statement

The authors declare no competing interests.

Acknowledgements

Gilles Vandewiele (1S31417N) and Isabelle Dehaene (1700520N) are funded by a scholarship of FWO. This study has been performed in the context of the 'Predictive health care using text analysis on unstructured data project', funded by imec, and the PRETURN (PREdiction Tool for prematUre laboR and Neonatal outcome) clinical trial (EC/2018/0609) of Ghent University Hospital. All funding bodies played no role in the creation of this study.

[1] `https://github.com/lusterck/preturn_ie`

Code and data availability

Code is available on Github under an open license. The datasets analyzed during the current study are not publicly available, due to privacy and security concerns, the underlying EHR data are not easy to distribute to researchers other than those engaged in the Institutional Review Board-approved research collaborations with the involved medical centers.

References

[1] P. Gooch, A modular, open-source information extraction framework for identifying clinical concepts and processes of care in clinical narratives (December 2012).

[2] T. B. Murdoch, A. S. Detsky, The inevitable application of big data to health care, Jama 309 (13) (2013) 1351–1352.

[3] W. Boag, D. Doss, T. Naumann, P. Szolovits, Whats in a note? unpacking predictive value in clinical note representations, AMIA Summits on Translational Science Proceedings 2018 (2018) 26.

[4] P. Resnik, M. Niv, M. Nossal, A. Kapit, R. Toren, Communication of clinically relevant information in electronic health records: a comparison between structured data and unrestricted physician language, Perspectives in Health Information Management.

[5] R. Grishman, Information extraction: Techniques and challenges, in: International summer school on information extraction, Springer, 1997, pp. 10–27.

[6] S. Chawanpaiboon, J. P. Vogel, A.-B. Moller, P. Lumbiganon, M. Petzold, D. Hogan, S. Landoulsi, N. Jampathong, K. Kongwattanakul, M. Laopaiboon, C. Lewis, S. Rattanakanokchai, D. N. Teng, J. Thinkhamrop, K. Watananirun, J. Zhang, W. Zhou, A. M. Glmezoglu, Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis, The Lancet Global Health 7 (1) (2019) e37 – e46. doi:https://doi.org/10.1016/S2214-109X(18)30451-0.
URL http://www.sciencedirect.com/science/article/pii/S2214109X18304510

[7] N. S. Wood, N. Marlow, K. Costeloe, A. T. Gibson, A. R. Wilkinson, Neurologic and developmental disability after extremely preterm birth, New England Journal of Medicine 343 (6) (2000) 378–384.

[8] R. L. Goldenberg, J. F. Culhane, J. D. Iams, R. Romero, Epidemiology and causes of preterm birth, The lancet 371 (9606) (2008) 75–84.

[9] R. K. Creasy, B. A. Gummer, G. C. Liggins, System for predicting spontaneous preterm birth., Obstetrics and Gynecology 55 (6) (1980) 692–695.

[10] G. Vandewiele, I. Dehaene, O. Janssens, F. Ongenae, F. De Backere, F. De Turck, K. Roelens, S. Van Hoecke, T. Demeester, Time-to-birth prediction models and the influence of expert opinions, in: D. Riaño, S. Wilk, A. ten Teije (Eds.), Artificial Intelligence in Medicine, Springer International Publishing, Cham, 2019, pp. 286–291.

[11] Y. Goldberg, J. Nivre, A dynamic oracle for arc-eager dependency parsing, in: COLING, 2012.

[12] N. Suff, L. Story, A. Shennan, The prediction of preterm delivery: What is new?, Seminars in Fetal and Neonatal Medicine 24 (1) (2019) 27 – 32, tHE CONTINUUM OF LATE PRETERM AND EARLY TERM BIRTHS. doi:https://doi.org/10.1016/j.siny.2018.09.006.
URL http://www.sciencedirect.com/science/article/pii/S1744165X18301112

[13] G. C. Liggins, R. N. Howie, et al., A controlled trial of antepartum glucocorticoid treatment for prevention of the respiratory distress syndrome in premature infants, Pediatrics 50 (4) (1972) 515–525.

[14] N. Melamed, J. Shah, A. Soraisham, E. W. Yoon, S. K. Lee, P. S. Shah, K. E. Murphy, Association between antenatal corticosteroid administration-to-birth interval and outcomes of preterm neonates, Obstetrics & Gynecology 125 (6) (2015) 1377–1384.

[15] S. Mehrabi, A. Krishnan, A. M. Roch, H. Schmidt, D. Li, J. Kesterson, C. Beesley, P. Dexter, M. Schmidt, M. Palakal, et al., Identification of patients with family history of pancreatic cancer-investigation of an nlp system portability, Studies in health technology and informatics 216 (2015) 604.

[16] Z. Tian, S. Sun, T. Eguale, C. M. Rochefort, Automated extraction of vte events from narrative radiology reports in electronic health records: a validation study, Medical care 55 (10) (2017) e73.

[17] G. K. Savova, J. Fan, Z. Ye, S. P. Murphy, J. Zheng, C. G. Chute, I. J. Kullo, Discovering peripheral arterial disease cases from radiology notes using natural language processing, in: AMIA Annual Symposium Proceedings, Vol. 2010, American Medical Informatics Association, 2010, p. 722.

[18] K. Jensen, C. Soguero-Ruiz, K. O. Mikalsen, R.-O. Lindsetmo, I. Kouskoumvekaki, M. Girolami, S. O. Skrovseth, K. M. Augestad, Analysis of free text in electronic health records for identification of cancer patient trajectories, Scientific reports 7 (2017) 46226.

[19] S. Sohn, G. K. Savova, Mayo clinic smoking status classification system: extensions and improvements, in: AMIA Annual Symposium Proceedings, Vol. 2009, American Medical Informatics Association, 2009, p. 619.

[20] J. Urbain, Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models, Journal of biomedical informatics 58 (2015) S143–S149.

[21] M. Topaz, L. Murga, K. M. Gaddis, M. V. McDonald, O. Bar-Bachar, Y. Goldberg, K. H. Bowles, Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches, Journal of biomedical informatics 90 (2019) 103103.

[22] J. Mantas, et al., An information extraction algorithm for detecting adverse events in neurosurgery using documents written in a natural rich-in-morphology language.

[23] H. Nassif, R. Woods, E. Burnside, M. Ayvaci, J. Shavlik, D. Page, Information extraction for clinical data mining: a mammography case study, in: 2009 IEEE International Conference on Data Mining Workshops, IEEE, 2009, pp. 37–42.

[24] A. Kormilitzin, N. Vaci, Q. Liu, A. Nevado-Holgado, Med7: a transferable clinical natural language processing model for electronic health records, arXiv preprint arXiv:2003.01271.

[25] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, H. Liu, Clinical information extraction applications: A literature review, Journal of Biomedical Informatics 77 (2018) 34 – 49. doi:https://doi.org/10.1016/j.jbi.2017.11.011.
URL http://www.sciencedirect.com/science/article/pii/S1532046417302563

[26] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, F. Puppe, Uima ruta: Rapid development of rule-based information extraction applications, Natural Language Engineering 22 (1) (2016) 1–40.

[27] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, to appear (2017).

[28] M. Neumann, D. King, I. Beltagy, W. Ammar, Scispacy: Fast and robust models for biomedical natural language processing, 2019. arXiv:arXiv:1902.07669.

[29] N. L. Andriy Mulyar, B. McInnes, Tac srie 2018: Extracting systematic review information with medacy, National Institute of Standards and Technology (NIST) 2018 Systematic Review Information Extraction (SRIE) ¿ Text Analysis Conference.

[30] L. J. Meertens, P. van Montfort, H. C. Scheepers, S. M. van Kuijk, R. Aardenburg, J. Langenveld, I. M. van Dooren, I. M. Zwaan, M. E. Spaanderman, L. J. Smits, Prediction models for the risk of spontaneous preterm birth based on maternal characteristics: a systematic review and independent external validation, Acta obstetricia et gynecologica Scandinavica.

[31] H. Watson, J. Carter, P. Seed, R. Tribe, A. Shennan, Quipp app: a safe alternative to a treat-all strategy for threatened preterm labor, Ultrasound in Obstetrics & Gynecology 50 (3) (2017) 342–346.

[32] D. A. De Silva, S. Lisonkova, P. von Dadelszen, A. R. Synnes, L. A. Magee, Timing of delivery in a high-risk obstetric population: a clinical prediction model, BMC pregnancy and childbirth 17 (1) (2017) 202.

[33] A. García-Blanco, V. Diago, V. S. De La Cruz, D. Hervás, C. Cháfer-Pericás, M. Vento, Can stress biomarkers predict preterm birth in women with threatened preterm labor?, Psychoneuroendocrinology 83 (2017) 19–24.

[34] G. Vandewiele, I. Dehaene, O. Janssens, F. Ongenae, F. De Backere, F. De Turck, K. Roelens, S. Van Hoecke, T. Demeester, Time-to-birth prediction models and the influence of expert opinions, in: Conference on Artificial Intelligence in Medicine in Europe, Springer, 2019, pp. 286–291.

[35] R. Leaman, R. Khare, Z. Lu, Challenges in clinical natural language processing for automated disorder normalization, Journal of Biomedical Informatics 57 (2015) 28 – 37. doi:https://doi.org/10.1016/j.jbi.2015.07.010.
URL http://www.sciencedirect.com/science/article/pii/S1532046415001501

[36] A. M. Cohen, W. R. Hersh, A survey of current work in biomedical text mining, Briefings in bioinformatics 6 (1) (2005) 57–71.

[37] P. Traung, The proposed new eu general data protection regulation, Computer Law Review International 13 (2) (2012) 33–49.

[38] S. Moen, T. S. S. Ananiadou, Distributional semantics resources for biomedical text processing, Proceedings of LBM (2013) 39–44.

[39] T. S. De Silva, D. MacDonald, G. Paterson, K. C. Sikdar, B. Cochrane, Systematized nomenclature of medicine clinical terms (snomed ct) to represent computed tomography procedures, Comput. Methods Prog. Biomed. 101 (3) (2011) 324329. doi:10.1016/j.cmpb.2011.01.002.
URL https://doi.org/10.1016/j.cmpb.2011.01.002

[40] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: Proceedings of NAACL-HLT, 2016, pp. 260–270.

[41] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (suppl_1) (2004) D267–D270.

[42] J. M. Zelle, R. J. Mooney, Learning to parse database queries using inductive logic programming, in: Proceedings of the national conference on artificial intelligence, 1996, pp. 1050–1055.

[43] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, B. G. Buchanan, A simple algorithm for identifying negated findings and diseases in discharge summaries, Journal of biomedical informatics 34 (5) (2001) 301–310.

[44] L. Sterckx, T. Demeester, J. Deleu, C. Develder, Knowledge base population using semantic label propagation, Knowledge-Based Systems 108 (2016) 79–91.

[45] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, Catboost: unbiased boosting with categorical features, in: Advances in Neural Information Processing Systems, 2018, pp. 6638–6648.

[46] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016, pp. 785–794.

[47] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, 2017, pp. 3146–3154.

[48] C. Click, M. Malohlava, A. Candel, H. Roark, V. Parmar, Gradient boosting machine with h2o, H2O. ai.

[49] A. V. Dorogush, V. Ershov, A. Gulin, Catboost: gradient boosting with categorical features support, arXiv preprint arXiv:1810.11363.

[50] A. Holzinger, C. Biemann, C. S. Pattichis, D. B. Kell, What do we need to build explainable ai systems for the medical domain?, arXiv preprint arXiv:1712.09923.

[51] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in neural information processing systems, 2017, pp. 4765–4774.

[52] S. M. Lundberg, G. G. Erion, S.-I. Lee, Consistent individualized feature attribution for tree ensembles, arXiv preprint arXiv:1802.03888.

[53] E. Celik, M. To, K. Gajewska, G. Smith, K. Nicolaides, Cervical length and obstetric history predict spontaneous preterm birth: development and validation of a model to provide individualized risk assessment, Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology 31 (5) (2008) 549–554.

[54] S. O'Hara, M. Zelesco, Z. Sun, Cervical length for predicting preterm birth and a comparison of ultrasonic measurement techniques, Australasian journal of ultrasound in medicine 16 (3) (2013) 124–134.

[55] J. Trienes, D. Trieschnigg, C. Seifert, D. Hiemstra, Comparing rule-based, feature-based and deep neural methods for de-identification of dutch medical records, arXiv preprint arXiv:2001.05714.

[56] D. Demner-Fushman, W. J. Rogers, A. R. Aronson, Metamap lite: an evaluation of a new java implementation of metamap, Journal of the American Medical Informatics Association : JAMIA 24 4 (2017) 841–844.

[57] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. K. Schuler, C. G. Chute, Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications, Journal of the American Medical Informatics Association : JAMIA 17 5 (2010) 507–13.