

Manuscript Number:

Title: A Framework for Extracting Information from Estonian Clinical Narratives

Article Type: Original Article

Keywords: Information Extraction; De-identification; Electronic Health Record; Clinical Decision Support; Estonian language

Corresponding Author: Mr. Raul Sirel,

Corresponding Author's Institution: Software Technology and Applications Competence Centre

First Author: Raul Sirel

Order of Authors: Raul Sirel; Timo Petmanson; Alexander Tkachenko; Dage Särg; Sven laur; Jaak Vilo

Abstract: Introduction. The number of clinicians using computers to manage their patient records has tremendously increased in the past two decades. However, as most of the data produced by the clinicians is unstructured, the applications of such data are limited. In order to combat the limitations of unstructured data, we propose a set of text-mining algorithms for extracting information from clinical narratives. Methods. In our proposed framework, the unstructured data is first de-identified by using CRF-based named entity recogniser. As no biomedical terminologies nor mapping software exist for Estonian, we developed a methods for creating dictionaries using clinical narratives and identifying dictionary elements in freetext. We also created grammars for extracting measurable observations (e.g. blood pressure) from freetext. Results. With the precision and recall respectively 95% and 97%, the performance of our de-identification system is comparable to other state of the art systems. Our dictionary creation and mapping algorithms yielded excellent precision (100%) and reasonable recall (72%). Our measurable observation extraction algorithm yielded precision and recall respectively 96% and 82%. Discussion. Our presented methods enable the usage of information otherwise locked away in clinical narratives. We argue that the extracted information can be utilised in both clinical decision support (CDS) and biomedical research. As an example application for CDS, we developed a system for creating visual reports on patient's health history.

Title:

A Framework for Extracting Information from Estonian Clinical Narratives

Address for Correspondence:

Raul Sirel

Postal Address: Software Technology and Applications Competence Centre, Ülikooli 2,
51003, Tartu, Estonia

E-mail: rsirel@ut.ee

Telephone: +372 56 150 661

Authors:

Raul SIREL, Software Technology and Applications Competence Centre, Tartu, Estonia;
Institute of Estonian and General Linguistics, University of Tartu, Tartu, Estonia

Timo PETMANSON, Software Technology and Applications Competence Centre, Tartu,
Estonia; Institute of Computer Science, University of Tartu, Tartu, Estonia

Alexander TKACHENKO, Software Technology and Applications Competence Centre,
Tartu, Estonia; Institute of Computer Science, University of Tartu, Tartu, Estonia

Dage SÄRG, Software Technology and Applications Competence Centre, Tartu, Estonia;
Institute of Estonian and General Linguistics, University of Tartu, Tartu, Estonia

Sven LAUR, Software Technology and Applications Competence Centre, Tartu, Estonia;
Institute of Computer Science, University of Tartu, Tartu, Estonia

Jaak VILO, Software Technology and Applications Competence Centre, Tartu, Estonia;
Institute of Computer Science, University of Tartu, Tartu, Estonia

Keywords:

Information Extraction; De-identification; Electronic Health Record; Clinical Decision
Support; Estonian language

ABSTRACT

Introduction. The number of clinicians using computers to manage their patient records has tremendously increased in the past two decades. However, as most of the data produced by the clinicians is unstructured, the applications of such data are limited. In order to combat the limitations of unstructured data, we propose a set of text-mining algorithms for extracting information from clinical narratives.

Methods. In our proposed framework, the unstructured data is first de-identified by using CRF-based named entity recogniser. As neither biomedical terminologies nor mapping software exist for Estonian, we developed methods for creating dictionaries using clinical narratives and identifying dictionary elements in freetext. We also created grammars for extracting measurable observations (e.g. blood pressure) from freetext.

Results. With the precision and recall respectively 95% and 97%, the performance of our de-identification system is comparable to other state of the art systems. Our dictionary creation and mapping algorithms yielded excellent precision (100%) and reasonable recall (72%). Our measurable observation extraction algorithm achieved precision and recall respectively 96% and 82%.

Discussion. Our presented methods enable the usage of information otherwise locked away in clinical narratives. We argue that the extracted information can be utilised in both clinical decision support (CDS) and biomedical research. As an example application for CDS, we developed a system for creating visual reports on patient's health history.

1. Introduction

Most of the clinicians nowadays have abandoned paper-based health records and are now using computers to manage their patient records (in 1998 already 76% of Estonian general practitioners (GPs) were using computers to manage their patients' health records [¹]). Such paradigm shift has resulted in enormous amounts of digital clinical data, commonly known as the electronic health record (EHR). Information in EHR holds a great promise from the perspective of clinical decision support (CDS) to improve patient safety and to support informed decision making [²].

However, to this day, the cornerstone of clinical documentation remains to be clinical narratives (admission notes, treatment plans, etc.) – according to Hicks [³], about 50% of the clinical data in EHR consists of narratives. Because of the heterogeneous nature of clinical narratives, the perspective of including freetext information from EHR to the clinical decision making process remains to be a grand challenge in CDS [⁴].

In order to make narrative-based CDS available for clinicians at the point of care, natural language processing (NLP) and text-mining are needed. As the development of such algorithms and models requires working with authentic data [⁵], it is vital to de-identify (i.e. to remove identifiers such as names, social security numbers, contact details, etc.) the data first. Although a number of papers have been published on de-identification of clinical texts, none of the proposed methods are applicable for Estonian in out-of-the-box manner. That is because most of the developed systems are based on statistical methods and thus require annotated corpora and additional lexical resources in the target language. No such resources existed for Estonian. To combat the problem, we created an annotated corpus of discharge letters and developed an automated system for de-identifying our EHR dataset using conditional random fields (CRF).

1 The integral component in extracting relevant clinical information from freetext data is
2 biomedical term identification (i.e., terminology mapping). That is, a process during which
3 biomedical terms (such as patient complaints, objective findings, drugs, etc.) present in text
4 are mapped onto some lexical resource (such as UMLS [⁶] or SNOMED CT [⁷]). In order to
5 successfully identify terms used in text, mapping software and dictionaries are needed. Even
6 though widely spoken languages, such as English, have multiple medical terminologies,
7 Estonian (and other small languages) is still terminologically rather ill-equipped.
8 Additionally, mapping solutions developed for English, such as MetaMap [⁸], are neither
9 applicable nor adaptable for Estonian or other morphologically complex languages.
10 Therefore, we devised a workflow for semi-automatic dictionary creation using a corpus of
11 EHR narratives and developed a robust mapping algorithm for identifying dictionary objects
12 in text using n-gram and skip-gram matching. Having such resources enabled us to create a
13 sample dictionary of intestinal complaints and successfully identify them in clinical
14 narratives.
15

16 EHRs also contain a variety of numerical measurements and observations such as
17 weight, height, blood pressure, blood sugar, etc. These basic measurements are often collected
18 during an examination and provide effective indicators to assess the general health of the
19 patient [⁹]. Additionally, combining such information with various laboratory results (e.g.
20 cholesterol levels) also facilitates early identification of obesity-related health risks linked to
21 diabetes and cardiovascular diseases [¹⁰, ¹¹]. Although our EHR dataset already contains
22 various information on measurements and observations (such as laboratory results) in a
23 structured form (in contrast to unstructured narratives), much of the information is still
24 presented in freetext (e.g. blood pressure measurements). Thus, we devised a method for
25 identifying sections of texts describing facts we call measurable observations by utilising
26 context-free grammars and regular expressions. Our method enabled us to model the
27

1 presentation of various numerical measurements and successfully identify them in clinical
2 narratives.
3

4 In this paper, we present a set of methods for extracting clinical information from
5 EHR narratives. We use NLP and text-mining to (a) preprocess and de-identify the patient
6 records; (b) identify clinical concepts despite not having a dictionary; (c) extract measurable
7 observations from clinical narratives. The novelty of our work resides in creating new
8 resources for processing and analysing Estonian clinical narratives and combining our
9 methods to enable clinicians to access the freetext information more easily. The research was
10 conducted for Estonian language by using the data of Estonian National Health Information
11 System (ENHIS) [¹²].
12
13
14
15
16
17
18
19
20
21
22
23
24
25

26 **2. Methods**

27 The data used in the described research was provided by the Estonian eHealth Foundation [¹³],
28 which administers the ENHIS. The ENHIS receives information from various Estonian
29 healthcare providers in the form of discharge reports. We were provided with discharge
30 reports sent to the ENHIS in 2012. The received dataset contained information about 561,793
31 patients in 1,371,486 discharge reports.
32
33
34
35
36
37
38
39
40

41 First, we preprocessed the narratives in our dataset by conducting tokenisation,
42 sentence splitting, and morphological analysis. Next, we used the data to develop a de-
43 identification system in order to de-identify the narratives prior to further research. Having
44 done that, we created formal grammars and dictionaries necessary for extracting measurable
45 observations and terms present in text. By doing that, we created grammars for several
46 measurable observations and a dictionary of patient complaints. Next, we identified
47 measurable observations and dictionary objects present in the de-identified ENHIS documents
48 and organised them in a database table. The overall workflow of our information extraction
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

framework is depicted on Figure 1 and it's key elements are described in the following chapters.

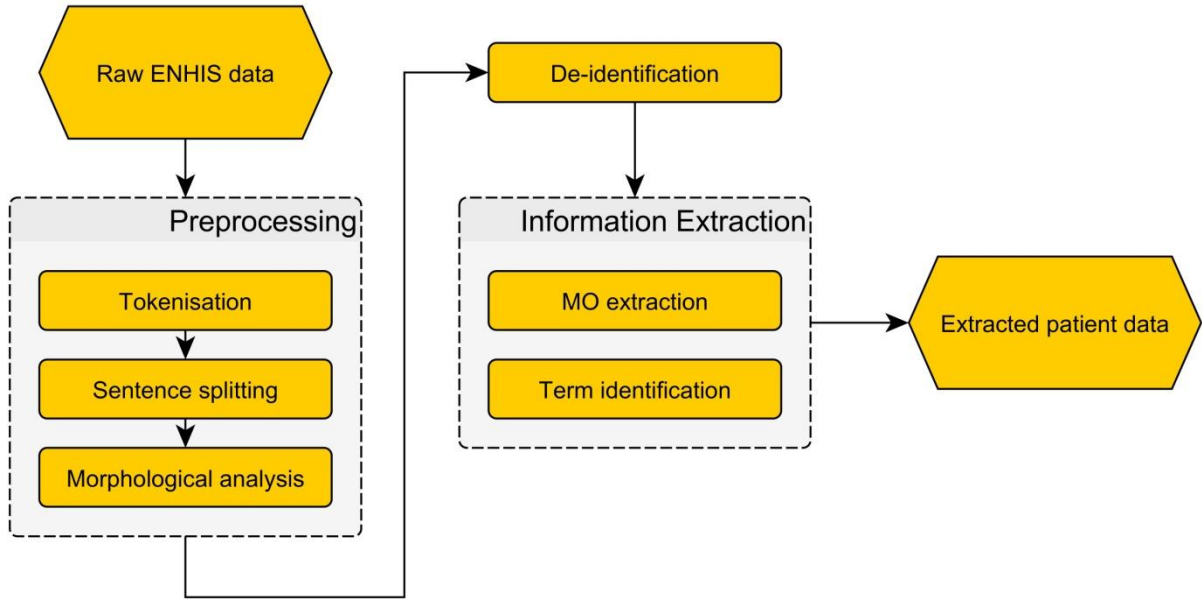


Figure 1. The general workflow of our information extraction framework; MO – measurable observation.

2.1. Preprocessing and de-identification

A common approach for de-identifying the EHR is named entity recognition (NER), a technique where entities like person names, locations, and organisations are automatically extracted from text. Various statistical and rule-based NER systems have been widely used in order to de-identify clinical texts written in English [14, 15, 16], Swedish [17] and French [18]. In the first i2b2 challenge, seven different de-identification systems were proposed [19]. One of the highest performing systems used the CRF algorithm, obtaining an F-score over 95%. Gardner et al. [20] used CRFs in their system HIDE, which achieved an overall accuracy of 98.2%. Another CRF-based system has demonstrated promising results in de-identifying Swedish clinical text [21].

A considerable issue in de-identification task is to decide which information can reveal person identity and therefore should be treated as protected health information (PHI). In fact,

several different ways of defining identifiable instances in EHR documents have been employed by different research groups [^{16, 17, 18, 19}]. In the United States, the HIPAA Act explicitly defines 16 PHI types, which need to be removed for clinical data to be considered de-identified [²²]. In Estonia, no similar regulations or guidelines are available [²³]. Therefore we defined a tag set covering 5 most frequently occurring PHI types: names, social security numbers, e-mail addresses, mobile and landline phone numbers.

For the purpose of person name extraction, we have adopted an existing NER system designed for entity recognition in Estonian news articles [²⁴]. It utilises the CRF learning algorithm and has a powerful toolkit for text preprocessing. We reused the existing pipeline to handle text tokenisation, sentence boundary detection, lemmatisation, part-of-speech (POS) tagging and feature extraction. In the pipeline, lemmatisation and POS-tagging were done by using morphological analyser [²⁵] and disambiguator [²⁶] first developed by FiloSoft Ltd [²⁷]. To adapt the system to a clinical domain, we compiled additional dictionaries for drugs, diseases and clinician names. Additionally, we defined a number of custom feature templates.

To train and evaluate our system, we have compiled a gold standard of 1855 discharge letters. To make sure that the dataset is representative, we selected records originating from different hospitals. The resulting corpus encompassed 27,131 sentences, 309,186 tokens. The vocabulary consisted of 43,259 unique words. In the corpus, all occurrences of person names have been tagged by a human annotator. We use a single tag to mark both first and second names of patients, relatives, clinicians and others. Overall 1,682 name instances were tagged. The corpus was then divided in development and test sets of 70% and 30% respectively. The development set was used for system tuning (finding relevant features and learning algorithm parameters), while the test set was reserved for the final evaluation run.

1 In contrast to person names, other PHI categories defined by us have more strict
2 structure and can be recognised by utilising a simple rule-based approach. For each category,
3
4 a set of regular expressions were implemented and all of the instances were replaced with a
5
6 corresponding tag.
7
8

9 As a result of preprocessing and de-identification, all narratives were tokenised,
10
11 sentence borders were identified, all words were enriched with their morphological
12
13 information including POS tags, and existing PHI was replaced with corresponding tags.
14
15
16
17
18

19 **2.2. Dictionary creation and mapping**

21 Specialist terminologies are often used in various NLP tasks in order to identify concepts in
22
23 biomedical or clinical texts. In order to combat the terminological deficiency, multiple
24
25 solutions have been introduced to automatically extract technical terms from domain corpora
26
27 [28, 29, 30, 31]. The main advantage of creating dictionaries by using domain corpora is the
28
29 perspective of capturing the actual language usage of the specialists. Similarly to some of the
30
31 methods proposed in the previous studies, we used a pattern-based approach to create
32
33 dictionaries containing clinical terms. Because of the extensiveness of the clinical domain, we
34
35 selected a subset of intestinal complaints to be developed first. We chose such a subset mainly
36
37 because we are conducting a study on colorectal cancer. Additionally, intestinal complaints
38
39 form a testset with a reasonable size to test the feasibility of the method.
40
41
42
43
44
45

46 We used the preprocessed narratives to create a corpus containing the anamneses of
47
48 the documents in ENHIS. Using the corpus (which also contained morphological information
49
50 from the preprocessing), we created collections of bi-, tri-, and quadgrams, to which we
51
52 applied Justeson-Katz filters [32] to select all possible term candidates. For example, we used
53
54 a filter to extract terms describing the presence of pain in various anatomical regions, which
55
56 required the presence of word *valu* (pain) followed by one or more nouns of which the last
57
58
59
60
61
62
63
64
65

had to be in the inessive case (locative grammatical case used in Finnish, Estonian, Basque, etc.): *valu.*_S_.*_S_.*in*. Alternatively, we used a filter to extract adpositional phrases containing the adposition *vahel* (between) which was followed by a noun in a nominative or partitive form: *_S_.*g,.*vahel+0.*_K_.*_S_.*[np]*.

The initial dictionary consisted of 2145 entries, but was afterwards manually revised by a linguist, which resulted in a dictionary containing 1278 noun phrases describing various intestinal complaints. The created dictionary also included numerous spelling variations and misspellings, which could later be used to capture more frequent variations of different terms. Using the created dictionary, our next step was to identify dictionary items present in the narratives of our ENHIS dataset. In order to do that, we devised a robust mapping algorithm using n-grams and skip-grams (Figure 2).

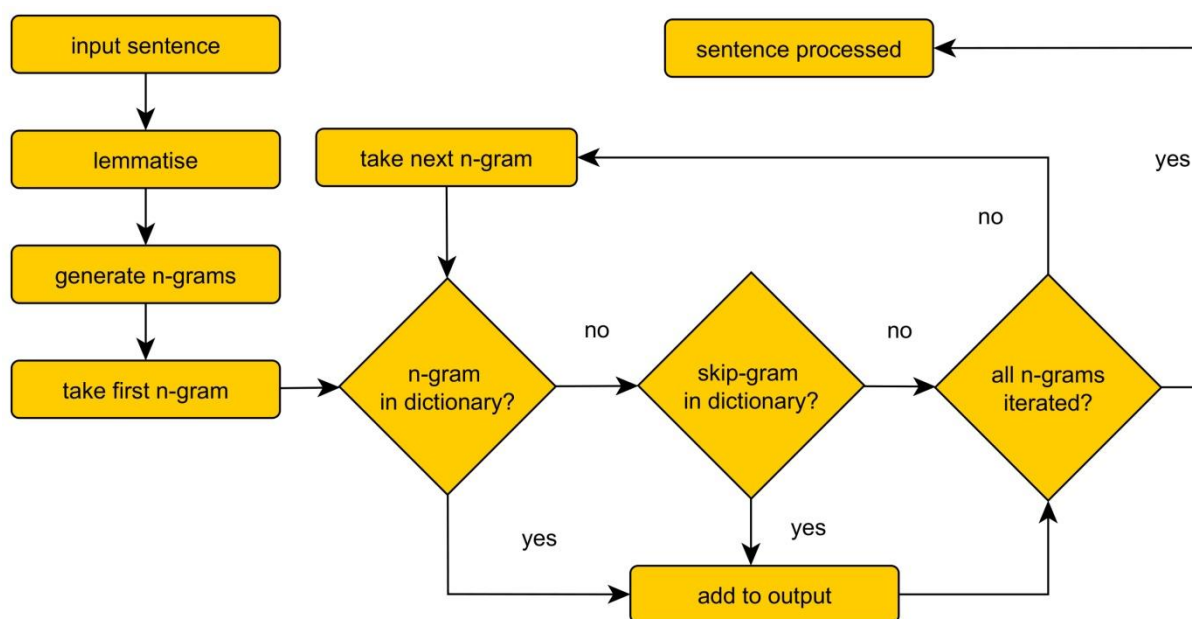


Figure 2. The mapping algorithm. We use skip-grams and dictionary permutation in order to compensate for the free word order of Estonian.

The algorithm of the mapper is following: first, the dictionary is imported and of every term in the dictionary, all possible permutations are generated. The permutations allow to

1 compensate for Estonian's relatively free word order. For example, both variants – *seljas valu*
2 and *valu seljas* (pain in back) – are grammatically correct and have the same meaning. As a
3
4 result of this process, an augmented dictionary is produced.
5
6

7 As an input for the mapper, we use the preprocessed and de-identified narratives. To match
8
9 the text against the augmented dictionary, we use n-gram and skip-gram matching. The latter
10
11 allows us to compensate for ellipsis (grammatical option to leave out words) and free word
12
13 order of Estonian.
14
15

16 From each sentence of the input, all possible n-grams are generated and lemmatised
17
18 (transformed into base forms). The generated lemma sequences of each n-gram is checked
19
20 against the augmented dictionary. If the lemma sequence is present in the augmented
21
22 dictionary, a corresponding term with its location and span is output from the original
23
24 dictionary. If the lemma sequence is not present in the augmented lexicon, a set of skip-grams
25
26 will be produced from the sequence. The skip-grams are generated by removing the lemmas
27
28 between the first and the last lemma of the n-gram, one at a time. For example, a noun phrase
29
30 *valu vasakus kõrvas* (pain in left ear) would produce a skip-gram *valu * kõrvas* (pain in * ear).
31
32 Then, each of the skip-grams will be checked against the augmented dictionary. If any of the
33
34 skip-grams is present in the dictionary, a corresponding term with its location and span is
35
36 output from the original dictionary.
37
38
39
40
41
42

43 While checking against the dictionary, we also use fuzzy matching in order to capture
44
45 less frequent spelling variations. For that, we allow Levenshtein distance [³³] of 1 between the
46
47 dictionary object and the n-gram. The distance of 1 was chosen by testing the balance of
48
49 precision and recall when increasing the allowed edit distance. The distance of 1 was the only
50
51 option where the recall was improved without impairing the precision.
52
53

54 After identifying a term in the text, it is tested for negation. Negation detection is an
55
56 integral part in any biomedical term mapping algorithm, for it is important to know whether
57
58
59
60
61
62
63
64
65

1 the term has a negative qualifier (for instance, one is not interested in cases where the
2 presence of a symptom has been negated). The negation detection algorithm utilised a rule-
3 based solution and the morphological annotations generated by the morphological analyser.
4 While in English, the negation is marked by a negating word (such as not, without, etc.) or a
5 negating preffix (such as a-, ab-, etc.), the Estonian negation markers, in addition to the ones
6 already mentioned, also include negating suffixes, abessive case, etc. For example, the
7 absence of pain may be expressed in various ways: *valutu*, *valuta*, *ilma valuta*, etc. The
8 algorithm tests the term by analysing it's morphological composition and local context (2
9 words before and after the term) for negation markers. The window of 2 was chosen by
10 observing the precision and recall when applying various windows sizes during the testing
11 phase.
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

29 **2.3. Extraction of measurable observations**

30
31 Although the ENHIS contains a wealth of information about measurable observations in a
32 structured manner (e.g. laboratory results), lot's of it is still presented among the clinical
33 narratives. For example, we determined that our ENHIS dataset did not contain any of the
34 following observations in a structured manner: blood pressure, pulse, weight, height, Apgar
35 score, etc. Additionally, we determined that even though data about biochemical
36 measurements such as creatinine, cholestrol, glycohemoglobin, etc. is available in a
37 structured manner, a substantial amount of these measurements were also presented as
38 freetext (Figure 3). As the accessibility of such information presented in freetext is rather
39 limited due to the sheer volume of the data, we devised a system for extracting measurable
40 observations from clinical narratives.
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

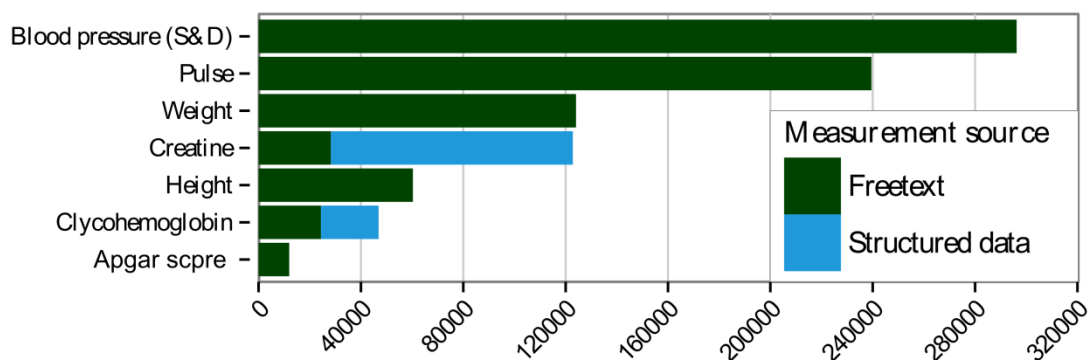


Figure 3. Combined measurement frequencies of blood pressure (systolic & diastolic), pulse, weight, creatinine, height, glycohemoglobin and APGAR score from freetext and structured datasets. Only certain set of measurement types are available in structured form.

When observing the data, we observed that measurable observations in clinical narratives have a rather strict structure. We saw that measurable observations in the clinical narratives are generally written in form of **OBS_NAME VALUE UNIT**. For example, in case of HDL-cholesterol (*high-density lipoprotein cholesterol*), **OBS_NAME** includes a range of variations such as *HDL*, *HDL-kol*, *HDL-kolesteroöl*, *hdl-Chol*. **VALUE** represents decimals covering variations written with comma, decimal point and optional spaces. For example, values like 2, 36 or 1.55 are both recognised. **UNIT** recognises variations like *mmol/L*, *mm/l* which are used to describe HDL-cholesterol in the texts.

In order to efficiently describe the name, value and unit parts for all the observation types, it is beneficial to write them as *production rules*, where the most generic rule covering every measurement is **OBSERVATION** \rightarrow **OBS_NAME VALUE UNIT**. In turn, we define **OBS_NAME** \rightarrow **HDL_NAME** | **LDL_NAME** | **CREATININE_NAME** | ... and describe the specifics of each observation in a separate production rule. For example, blood pressure measurements like *120 / 90 mmHg* have separate systolic and diastolic values, which can be captured by defining **BP_VALUE** \rightarrow **SYSTOLIC / DIASTOLIC**. As a result, variations of

different observations types are described in a maintainable and easily extendable way. An example of our production rules can be seen below:

```
OBSERVATION → OBS_NAME VALUE UNIT
OBS_NAME → HDL_NAME | LDL_NAME | BP_NAME | ...
VALUE → BP_VALUE | DECIMAL_VALUE
UNIT → HDL_UNIT | LDL_UNIT | BP_UNIT | ...
BP_VALUE → SYSTOLIC / DIASTOLIC
SYSTOLIC → DECIMAL_VALUE
DIASTOLIC → DECIMAL_VALUE
HDL_NAME → HDL kolesterool | HDL-kolesterool | ...
HDL_UNIT → mmol/L | mm/l
```

The production rules described above actually make up a context-free grammar (CFG), first proposed by Noam Chomsky [³⁴]. Context-free grammars and their variants, such as probabilistic context-free grammars (PCFG), have proven to be successful in numerous information extraction systems in the past, including clinical text-mining for extracting medication data from clinical notes [³⁵, ³⁶, ³⁷, ³⁸].

The creation of our CFG was a laborious task – the ENHIS dataset contains many variations of how observations can be written: variations in names and units, variations in the order of the tokens e.g. *hdl (mmol/L) 3,5*, extra words e.g. *hdl oli 3,5* (hdl was 3.5), etc. The main technique we used to facilitate our work was automatic generalisation of parts of our grammar. For example, given a string *hdl-kolesterool*, that describes HDL-cholesterol, we automatically generated regular expressions such as *hdl[a-z]** (matching any word starting with *hdl*). By replacing the original version, we allowed more name variations and vastly increased the coverage of our matches. However, we also increased the risk of obtaining false positive matches. Therefore, we automatically evaluated the precision for each generalised

1 candidate on a manually annotated subset of the dataset and chose best candidates for manual
2 verification. After that step, we decided whether to include or exclude a particular candidate.
3
4 During this process, we were also able to incrementally increase the size of the annotated
5 dataset used for precision evaluation. Another form of generalisation was making parts of the
6
7 grammar optional.
8
9

10
11 The grammar we constructed is ambiguous, i.e. there are often more than one way to
12 parse the input string. For example, given a phrase *hdl kolesterool 2 kuu pärast kõrgem*
13 (HDL-cholesterol higher after 2 months), we match *kolesterool 2*, we also match *hdl*
14 *kolesterool 2*, and finally, we also have a time expression (timex) *2 kuu pärast*. One could
15
16 assume that the conflict could be solved by choosing the longer match. However, both of the
17 longer matches would be incorrect and instead we should choose the timex. In order to solve
18
19 this issue, we assign each match a weight, that defaults to its length, and a coefficient, which
20 defaults to one and depends on parts of the grammar that was used to parse the match. We
21 manually adjust the coefficients by analysing the incorrect matches we find. This is only
22 required for special cases, such as the timexes, as default coefficients usually yield correct
23 results.
24
25

26
27 As previously noted, there are observations without units in the dataset. They are
28 usually omitted by the clinicians, when it is easy to infer the unit from the order of the
29 magnitude of the value. For example, phrase *the height is 167* can be assumed to have
30 centimeter unit while *the height is 1.67* is in meters. As the unit is crucial for epidemiological
31 studies, we use support vector machine [39] models to predict the units for the missing
32 examples. This works with high precision very close or equal to 100% for every observation
33 type.
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

3. Results

3.1. De-identification

The final de-identification system demonstrated the precision and recall of respectively 95% and 97%. Compared to a baseline system, which simply matches entities contained in a dictionary of common person names, this translates into 45pp improvement in F-score (Table 1). It also appears that the extraction of person names in clinical texts is somewhat easier than in the news domain, where the most recent result for Estonian is 89% in both precision and recall [23]. Overall, our de-identification system's performance is comparable to the state of the art results reported by other authors [19] (Table 1).

	Precision, %	Recall, %	F-score, %
Baseline system	40.0	70.0	51.0
CRF-based system	95.0	97.0	96.0
Szarvas et al.	97.8	95.7	96.7
Wellner et al.	97.0	95.9	96.3

Table 1. Evaluation results of our de-identification system.

3.2. Dictionary creation and mapping

As defining a gold standard for a mapping experiment without having massive biomedical terminologies is a difficult task, we decided to evaluate the dictionary creation and mapping methods in a combined mapping experiment. First, we combined a subset of 500 narratives written about patients with intestinal diseases in a corpus and mapped it onto the dictionary of intestinal complaints. We then manually checked the narratives and their mappings to determine the amount of incorrect mappings, and of terms referring to abdominal discomfort/pain appearing in freetext, but missing in mappings (or missing in the dictionary altogether). We concluded that our system performed with excellent precision (we found no

false positive mappings) and with moderate recall (72%) as not all of the variations were captured during the dictionary creation process.

Additionally, we performed an experiment, where we disabled several functionalities of the mapper (permutations, skip-gram matching, Levenshtein distance) one at a time in order to see it's effect on recall. The aim of the experiment was to determine the effect of the mapping functionalities on the mapping quality. The most significant drop was produced by removing the step of permutating dictionaries from the algorithm. The drop was somewhat lower when removing skip-gram creation and fuzzy matching (Table 2).

Removed feature	Recall, %	Drop, %
None (baseline)	71.8	0.0
Fuzzy matching	70.2	1.7
Skip-gram	67.4	4.4
Permutations	64.1	7.7

Table 2. Results of the leave-one-out experiment.

3.3. Extraction of measurable observations

Using our approach for modelling measurable observations, we created a CFG for 23 measurable observations and extracted a total of 1,031,253 measurable observations in our ENHIS dataset. Combined with structured data of 164,270 measurements, we got total of 1,195,582 data points, where 59 were duplicates (same patient, same measurement and value for same date). Majority of measurements (86 percent) come from free-text (Figure 3).

In order to evaluate the precision and recall of our grammar, we compiled test sets of discharge letters based on the ICD-10 codes assigned to the documents. The first set contained regular health examination discharges (ICD-10 codes Z00-Z13, Z30-Z39). The second set contained the essential hypertension discharges (ICD-10 code I10), and the last group

1 contained *diabetes mellitus* discharges (ICD-10 codes E10-E14). Then, we drew a random
2 sample of 300 discharges from each set and manually annotated all true positive, false
3 positive and false negative matches, a total of 3099 annotations. The estimated overall
4 precision was 96% and overall recall was 82%. The detailed evaluation results for each
5 observation type are given in Appendix I.
6
7
8
9
10

11 We chose three different groups instead of one in order to capture more variations in
12 the structure of the measurable observations. Also, we expected these particular groups to
13 contain observations our grammar was designed to recognise.
14
15
16
17
18
19
20
21

22 **4. Discussion**

23 In this paper, we proposed a set of methods and algorithms for increasing the value of
24 information locked away as clinical narratives. Using the ENHIS dataset, we first identified
25 six critical information types that needed to be removed for the clinical data to be considered
26 de-identified. In order to remove such information from the narratives, we built a de-
27 identification system using the CRF machine learning algorithm. Having the data de-
28 identified, we developed a pipeline for creating clinical dictionaries and extracting dictionary
29 objects from clinical narratives, and a system for defining and extracting measurable
30 observations from freetext.
31
32
33
34
35
36
37
38
39
40
41
42

43 The extracted information holds a great potential as it has multiple applications. For
44 example, we used it in a software prototype for generating visual reports on patient's health
45 history. Our prototype allows us to visualise various measurable observations and intestinal
46 complaints described in freetext during earlier patient-doctor encounters (Figure 4). Such
47 visualisations enable easy access to information otherwise likely disregarded because of the
48 limited time allocated to each patient-doctor encounter.
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

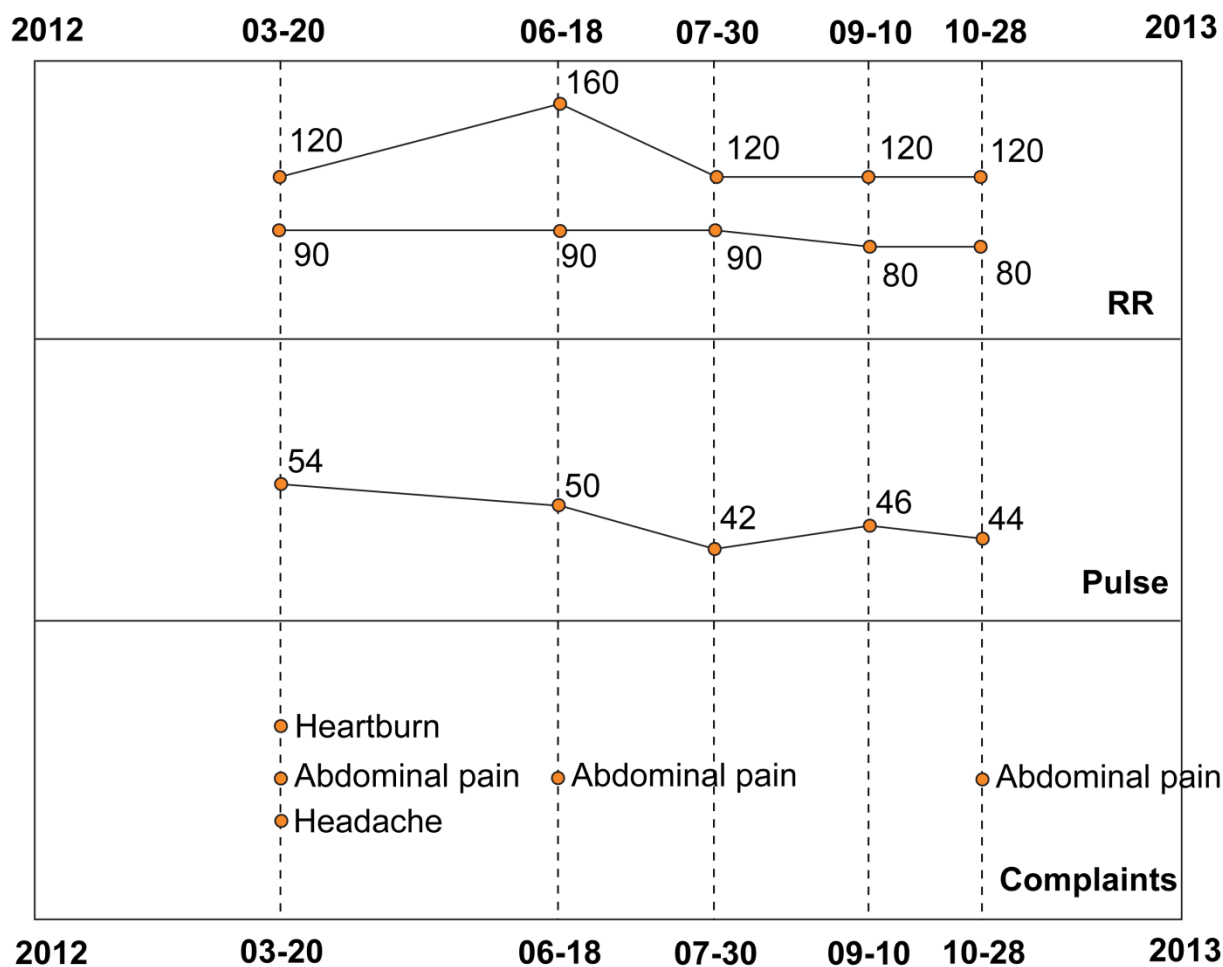


Figure 4. An example application for the information extracted by using our methods – a temporal visualisation of patient's history. Each vertical line depicts a doctor-patient encounter, during which information is inserted.

Additionally, the extracted information can also be used in biomedical research. For example, we have extracted the PSA (prostate-specific antigen) levels described in freetext and are currently using it (combined with other information) for building diagnostic models for prostate cancer. We are also using the extracted intestinal complaints in order to create diagnostic models for colorectal cancer. Furthermore, extracted blood pressure measurements combined with prescription data can also be used in researching the development of drug resistance in hypertensive patients.

1 In future work, we plan to further investigate other entities which occurrence or co-
2 occurrence may still reveal person's identity – our aim is to define a wider set of PHI types.
3
4 We're also planning to further refine our information extraction algorithms in order to yield
5 higher recall, and define new dictionaries and measurable observation grammars to increase
6 the volume of extracted information.
7
8
9
10

11 In conclusion, we find that clinical narratives contain a wealth of information that may
12 be utilised in both CDS and biomedical research. Despite the fact that standard information
13 extraction systems do not support small languages such as Estonian, we have built a
14 framework using simple methods that yield reasonable results. Our main contributions are the
15 developed methods and algorithms for preprocessing, de-identifying, and extracting
16 information from clinical narratives, as no such resources existed for Estonian language
17 before. Although our methods were developed for Estonian, the general concept is applicable
18 to other languages as well, especially those with rich morphology and insufficient
19 terminological resources.
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

ACKNOWLEDGEMENTS

The research was conducted in the Software Technology and Applications Competence Centre (STACC). STACC is a research and development organisation funded from the European Regional Development Fund. The data used in the research was provided by the Estonian eHealth Foundation.

APPENDIX I: Precision and recall of extracted observations

Type	N	Precision (%)	Recall (%)
Blood pressure	728	93 (91 .. 95)	96 (94 .. 97)
Pulse	630	98 (97 .. 99)	77 (73 .. 80)
Weight	300	96 (92 .. 97.6)	75 (70 .. 80)
Blood sugar	266	100 (98 .. 100)	57 (51 .. 63)
Hemoglobin	166	95 (90 .. 97.4)	96 (92 .. 98)
Creatinine	146	100 (97 .. 100)	93 (88 .. 96)
Height	134	94 (87 .. 96.9)	81 (73 .. 87)
Glycohemoglobin	129	100 (97 .. 100)	87 (80 .. 92)
Free cholesterol	102	99 (94 .. 99.8)	90 (83 .. 95)
Body mass index	95	100 (94 .. 100)	66 (56 .. 75)
Aspartate aminotransferase	57	100 (94 .. 100)	97 (88 .. 99)
Alanine aminotransferase	55	100 (93 .. 100)	96 (88 .. 99)
LDL-cholesterol	48	100 (92 .. 100)	88 (75 .. 94)
HDL-cholesterol	46	100 (91 .. 100)	83 (69 .. 91)
Uric acid	39	100 (91 .. 100)	95 (83 .. 99)
Triglycerides	34	100 (88 .. 100)	82 (67 .. 92)
Wound length	34	71 (54 .. 83.2)	100 (86 .. 100)
Apgar score	29	100 (80 .. 100)	52 (34 .. 67)
Prostate-Specific Antigen	23	100 (85 .. 100)	96 (79 .. 99)
Prostate size	17	94 (72 .. 98.9)	94 (72 .. 99)
Waist circumference	11	100 (51 .. 100)	36 (15 .. 65)
Prostate biopsy	8	100 (51 .. 100)	50 (22 .. 79)

Testosterone	2	100 (34 .. 100)	100 (34 .. 100)
TOTAL	3099	96 (96 .. 97)	82 (82 .. 84)

Table 3. Precision and recall estimates of extracted observations on the observation extraction testing dataset, containing a total of 900 discharges. The 95% confidence intervals are estimated using Wilson score interval.

REFERENCES

- 1 Kalda R, Lember M. Setting national standards for practice equipment. Presence of
equipment in Estonian practices before and after introduction of guidelines with
feedback. *International Journal for Quality in Health Care* 2000; 12(1): 59-63.
- 2 Jensen JB, Jensen LJ, Brunak S. Mining electronic health records: towards better
research applications and clinical care. *Nature Reviews Genetics* 2012; 13: 395-405.
- 3 Hicks J. The potential of claims data to support the measurement of health care
quality. San Diego, CA: RAND; 2003.
- 4 Sittig DF, Ash JS, Ledley RS. The Story Behind the Development of the First Whole-
body Computerized Tomography Scanner as Told by Robert S. Ledley. *J Am Med
Inform Assoc.* 2006 Sep-Oct; 13(5): 465-469.
- 5 Kilgarriff A. Comparing Corpora. *International Journal of Corpus Linguistics* 2001; 6
(1), 1-37.
- 6 Unified Medical Language System (UMLS). The UMLS Page. Available at:
<http://www.nlm.nih.gov/research/umls/>. Accessed September 18, 2014.
- 7 SNOMED CT. The IHTSDO Page. Available at: <http://www.ihtsdo.org/snomed-ct/>.
Accessed September 18, 2014.
- 8 Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent
advances. *J Am Med Inform Assoc.* 2010 May-Jun; 17(3): 229-36.
- 9 Mokdad AH, Ford ES, Bowman BA, et al. Prevalence of obesity, diabetes, and
obesity-related health risk factors. *Jama* 2003; 289(1): 76-79.
- 10 Chan JM, Rimm EB, Colditz GA, et al. Obesity, fat distribution, and weight gain as
risk factors for clinical diabetes in men. *Diabetes care* 1994; 17(9): 961-969.
- 11 Mackay J, Mensah GA. The atlas of heart disease and stroke 2004. Available at:
http://www.who.int/cardiovascular_diseases/resources/atlas/en/. Accessed September
18, 2014.
- 12 Estonian National Health Information System. The Estonia eHealth Foundation Page.
Available at: <http://www.e-tervis.ee/index.php/en/health-information-system>.
Accessed September 18, 2014.
- 13 Estonian eHealth Foundation. The Estonian eHealth Foundation Page. Available at:
<http://www.e-tervis.ee/index.php/en/>. Accessed September 18, 2014.
- 14 Uzuner Ö, Sibanda TC, Luo Y, Szolovits P: A De-identifier for Medical Discharge
Summaries. *Journal of Artificial Intelligence in Medicine* 2008, 42(1):13-35.

-
- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
- 15 Sweeney L. Replacing Personally-Identifying Information in Medical Records, the Scrub System. Proceedings of The AMIA Annual. Fall Symposium 1996:333-337.
- 16 Neamatullah IM, Douglass M, Lehman LH, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD: Automated De-identification of Free Text Medical Records. BMC Medical Informatics and Decision Making 2008, 8:32.
- 17 Kokkinakis D, Thurin A: Identification of Entity References in Hospital Discharge Letters. In Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007 University of Tartu, Tartu; 2007.
- 18 Grouin C, Rosier R, Dameron O, Zweigenbaum P: Testing Tactics to localize de-identification. Studies in health technology and informatics 2009, 150:735-739.
- 19 Uzuner Ö, Luo Y, Szolovits P: Evaluating the State-of-the-art in Automatic De-identification. Journal of the American Medical Informatics Association 2007, 14(5):550-563.
- 20
- 21 Dalianis H, Velupillai S. De-identifying Swedish clinical text-refinement of a gold standard and experiments with Conditional random fields. J. Biomedical Semantics 1 (2010): 6.
- 22 HIPAA: Health Insurance Portability and Accountability (HIPAA), Privacy Rule and Public Health Guidance. From CD Cand the U.S. Department of Health and Human Services 2003. Available at: <http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm>. Accessed September 18, 2014.
- 23 Personal Data Protection Act. The Riigi Teataja Page. Available at: <https://www.riigiteataja.ee/en/eli/509072014018/consolide>. Accessed at September 18, 2014.
- 24 Tkachenko A, Petmanson T, Laur S. 2013. Named Entity Recognition in Estonian. ACL 2013: 78.
- 25 Kaalep, HJ 1997. An Estonian morphological analyser and the impact of a corpus on its development. – Computers and the Humanities, 31 (2), 115–133.
- 26 Kaalep HJ, Vaino T. Complete morphological analysis in the linguist's toolbox. In: Proceedings of Congressus Nonus Internationalis Fenno-Ugristarum 2001; 9-16.
- 27 FiloSoft Ltd. The FiloSoft Page. Available at: <http://www.filosoft.ee>. Accessed September 18, 2014.
- 28 Frantzi K, Ananiadou S, Tsujii J. The C-value/NC-value Method of Automatic Recognition of Multi-word Terms, In: ECDL '98 Proceedings of the Second European

-
- Conference on Research and Advanced Technology for Digital Libraries 1998; 585-604.
- 29 Navigli R, Velardi P. Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics* 2004. 30 (2): 151-179.
- 30 Park Y, Byrd RJ, Boguraev B. Automatic glossary extraction: beyond terminology identification. *Proceedings of International Conference On Computational Linguistics* 2002; 1-7.
- 31 Wermter J, Hahn U. Finding New terminology in Very large Corpora. *Proceedings of the 3rd International Conference on Knowledge Capture* 2005. Available at: http://pdf.aminer.org/000/472/249/finding_new_terminology_in_very_large_corpora.pdf. Accessed September 18, 2014.
- 32 Justeson S, Katz S. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1995; 1(1): 9-27.
- 33 Влади́мир И. Левенштейн. Двоичные коды с исправлением выпадений, вставок и замещений символов [Binary codes capable of correcting deletions, insertions, and reversals]. Доклады Академий Наук СССР (in Russian) 163 (4): 845–848. Appeared in English as: Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 1966; 10(8): 707–710.
- 34 Hopcroft JE.; Ullman JD. 1979. *Introduction to Automata Theory, Languages, and Computation* (1st ed.). Addison-Wesley. Pp 106.
- 35 Temkin JM, Gilder MR. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics* 2003; 19(16): 2046-2053.
- 36 Huck G, Frankhauser P, Aberer K et al. Jedi: Extracting and synthesizing information from the web. In: *Cooperative Information Systems* 1998: 32-41.
- 37 Viola P, Narasimhan M. Learning to extract information from semi-structured text using a discriminative context free grammar. In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* 2005; 330-337.
- 38 Xu H, Stenner SP, Doan S et al. MedEx: a medication information extraction system for clinical narratives. *JAMIA* 2010; 17(1): 19-24.
- 39 Cortes C, Vapnik V. Support-vector networks. *Machine learning* 1995; 20(3): 273-297.

SUMMARY TABLE

What was already known on the topic:

- Natural language processing has been successfully used for extracting information from clinical narratives
- Existing NLP resources developed for English are not applicable for other languages
- For de-identification, CRF-based solutions have outperformed the alternatives
- Context free grammars have been successfully deployed for medical information extraction on non-Estonian corpora.

What this study added to our knowledge:

- A CRF-based de-identification solution performed adequately for Estonian clinical narratives
- It is feasible to identify clinical terms despite not having extensive clinical terminologies
- Using context free grammars for extracting measurable observation gives acceptable precision and recall on the ENHIS dataset.

CONTRIBUTOR STATEMENT

The de-identification part of the study was conducted by Alexander Tkachenko. The dictionary creation and mapping part was done by Raul Sirel, supported by Dage Särg. The numerical measurement extraction part was conducted by Timo Petmanson, supported by Dage Särg. The experiments described in the results section were designed and conducted by Raul Sirel, Timo Petmanson, Alexander Tkachenko, and Dage Särg. Sven Laur contributed to this study with his supervision and critical feedback. The manuscript was written by Raul Sirel, Timo Petmanson, and Alexander Tkachenko, the remaining authors supported the writing process with their critical comments and suggestions. Prof. Jaak Vilo was responsible for providing resources and infrastructure necessary for the research.

Raul Sirel

Timo Petmanson

Alexander Tkachenko

Dage Särg

Sven Laur

Jaak Vilo

CONFLICT OF INTEREST DECLARATION

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

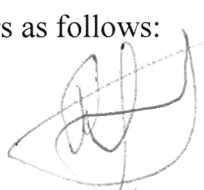
We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

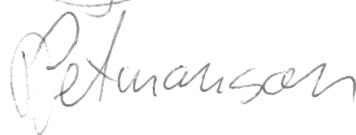
We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author.

Signed by all authors as follows:

Raul Sirel



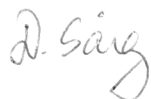
Timo Petmanson



Alexander Tkachenko



Dage Särg



Jaak Vilo



SVEN LAUR

