Similarity Analyzer for Semantic Interoperability of Electronic Health Records Using Artificial Intelligence (AI)

A Naveed[1*], Y F Hu[1], T Sigwele[2], G Mohi-Ud-Din[4], M Susanto[3], and M Ali[1]

[1]*Faculty of Engineering and Informatics, University of Bradford, BD7 1DP, United Kingdom,*

[2]*Department of Computing and Informatics, Faculty of Science, BIUST University, Private Bag 16, Palapye, Botswana,*

[3]*Department of Electrical Engineering, Faculty of Engineering, University of Lampung, Jl. Prof. Sumantri Brojonegoro No. 1, Bandar Lampung 35145, Indonesia*

[4]*Department of Computer Sciences, Faculty of Science, Liverpool John Moores University, Liverpool, United Kingdom*

[*]*Email: a.naveed2@bradford.ac.uk*

Article Information

Received:
30 September 2019

Received in revised form:
27 October 2019

Accepted:
14 November 2019

Volume 1, Issue 2, December 2019
pp. 53 – 58

http://dx.doi.org/10.23960/jesr.v1i2.13

Abstract

The introduction of Electronic Health Records (EHR) has opened possibilities for solving interoperability issues within the healthcare sector. However, even with the introduction of EHRs, healthcare systems like hospitals and pharmacies remain isolated with no sharing of EHRs due to semantic interoperability issues. This paper extends our previous work in which we proposed a framework that dealt with semantic interoperability and security of EHR. The extension is the proposal of a cloud-based similarity analyzer for data structuring, data mapping, data modeling and conflict removal using Word2vec Artificial Intelligence (AI) technique. Different types of conflicts are removed from data in order to model data into common data types which can be interpreted by different stakeholders.

Keywords: *semantic interoperability; interoperability standards;electronic health records (EHR);artifical intelligence techniques.*

I. INTRODUCTION

Recently, healthcare organizations have gradually migrated paper-based patient medical records to digital electronic ones by the implementation of Electronic Health Records (EHR) systems which is a paradigm shift in the healthcare sector [1]. Various EHR standards exists like IEEE DICOM, LOINC, SNOMED CT [2], HL7 and FHIR [3]. However, even with the introduction of EHR and its diver's standards, healthcare systems are still isolated from each other with no collaboration and interoperability.

Interoperability is the ability of two or more components, applications or systems to exchange and use information. Interoperability of EHR defined in Health Information Management System Society (HIMSS) as **"**the ability of two or more applications being able to communicate in an effective manner without compromising the contents of transmitted EHR**"** [3]**.** The data of EHR can be shared within different units of hospitals (intra-sharing) or between different units (inter-sharing), between different laboratories and external agencies such as insurance and other research units as shown in Fig. 1 [4].

The major goal of interoperability in healthcare is to facilitate the seamless exchange of healthcare related data and an environment is needed which supports interoperability and secures transfer of data. Healthcare Interoperability has the following advantages: easy access of patient's records; reduction of medical errors hence less casualties; healthcare cost reduction and reducing delays in medal healthcare systems.

Some of the issues that require our attention to achieving complete interoperability of shareable EHR systems are as follow: partial mapping from multiple sources [1]; need of user intervention; setting of standards/guideline; addressing contextual constraints; existence of semantic differences in attributes;

platforms for semantic interoperability; ontology mapping [4]; and interpreting medical terminologies [5,6].

In the context of interoperability, the key security issues are: whom to share; how to share; where to share that EHR data with such that no unauthorized access can be made to any data [7]. Another important challenge is assignment of authorization and access of required data to authorized person [8]. Moreover, ensuring confidentiality and privacy of patient's sensitive health data shared within the departments of one hospital as well as between different hospitals is another challenge to be addressed [9,10].

This paper proposed a framework that addresses both the interoperability and security issues in electronic health records in our previous paper [1]. Also, it extends our previous work [1]. This mainly focused on the third layer of our proposed model which deals with the semantic interoperability of Electronic Health Records (EHR) using Artificial Intelligence (AI) Techniques.

II. MATERIALS AND METHODS

A. Related Work

Extensive research has been done on semantic interoperability of electronic health records. Authors in [8] explained that achieving semantic interoperability requires user intervention and thus limits the possibility of controlling and managing secured sharing of EHRs dynamically. Syntactic interoperability on the other hand has low-level technical issues like that of formats, schema and protocols that can be resolved using various techniques and approaches. Semantic interoperability requires different levels of integration in inter as well as intra organizations and is difficult to obtain.

Also, it is observed that healthcare domain exhibits data having high sensitivity in terms of required security. Moreover, the need of EHR security differs from person to person or case to case. Hence, a dynamic and robust technique or approach must be appropriately selected for permitting secured sharing of sensitive health data in disparate interoperable healthcare domain. Authors in [10], developed a model which is based on ontology for interoperability between heterogeneous systems. The authors focus on modeling, structuring, representing data along with its interoperability.

There are various ways to model and represent data such as SNOMED, however, they lack in providing full interoperability. The approaches such as knowledge base and ontology frameworks are widely adopted for providing full interoperability. The UntolUrgences is an ontology-based framework for the emergency acts. Another ontology-based framework is proposed to model medical decision support system to improve patient's lifestyle. Other paper described that EHR solutions are complex, spanning multiple specialties and domains of expertise [11]. These systems need to handle clinical concepts, temporal data, documents, and financial transactions, which leads to a large code base that is tightly coupled with data models and inherently hard to maintain.

These difficulties can greatly increase the cost of developing EHR systems, result in a high failure rate of implementation, and threaten investments in this sector. Moreover, due to the wide variance in the level of detail across different settings, data exchange is becoming a serious problem, further increasing the cost of development and maintenance. Others stated that semantic interoperability is of prime importance for healthcare systems to communicate with each other and provide better healthcare facilities to patients [12].

Compatibility between heterogeneous healthcare standards for message schemas conversions requires ontology matching tools. The proposed system uses ontology matching tools to resolve the data level heterogeneities between different healthcare standards and achieve message schema level conversion. Services based on ontology matching helps healthcare systems to communicate with any other system. Therefore, in future main focus will be on working towards establishing more accurate mapping services and more detail level interaction study of existing healthcare Standards mapping services based on Surface Oriented Architecture (SOA).

It also explained that semantic interoperability challenges [13]. They explained the variety and veracity dimensions for data analysis and decision making applications in healthcare data. Many issues are raised while dealing with interoperability mainly with standards. They discussed that for improvement of information sharing and addressing the problems of data medication with domain ontologies, semantics play an important role. They then explained the main steps for building the domain ontologies for Forensic and Legal medicine. They concluded that ontologies can be used to enrich data and to query data stored in large heterogeneous databases.

B. Proposed Interoperability Framework with Similarity Analyser

1. Detailed Framework Description

A framework that deals both with the semantic interoperability of EHR is proposed in our previous paper [1].Our proposed framework is divided into 4 layers as shown in Fig. 1.

Layer 1- Data layer: The first layer manages data in the cloud. This layer contains repositories to store data related to EHR from hospitals. All information in documents like patient information, EHR's and other system of records located on cloud will be stored here. On this layer, MySQL database is used to store data.

Layer 2- Syntactic Interoperability Layer: This layer will define all the archetypes related to the different kinds of data such as blood pressure and Syntactic separation of the EHR data. This means that data is extracted from the database from first layer and separated into various sub categories such as clinical, personal, and financial and research related data into meaningful entities. Fast Healthcare Interoperability Resource (FHIR) is used here.

Layer 3- Semantic Interoperability Layer: This layer will define all the repositories to store archetypes and is responsible for semantic interoperability of the EHR dataset. This layer is divided into two sub categories, model of use and model of meaning. Model of use include generic information model and data structure of healthcare data. Model of meaning include different health terminologies and for this we will use SNOMED CT standard and domain level and top level ontology will be treated here.

For semantic interoperability the similarity analyzer is very important and is placed with the cloud based EHR. Similarity analyzer performs various functions such as data structuring, data mapping, data modeling and conflict removal. Data is structured into various archetypes which provide specific information about an object such as blood pressure. Different types of conflicts are removed from the data to model data into common types which can be interpreted by different stakeholders. The similarity analyzer is fully explained in the part B of this section.

Layer 4: Data Exchange Layer: This layer defines how the data will be transferred to different stakeholders. Archetypes specify the design of the clinical data that a Health Care Professional needs to store in a computer system. Archetypes enable the formal definition of clinical content by domain experts without the need for technical understanding. These conserve the meaning of data by maintaining explicitly specified and well-structured clinical content for semantic interoperability. These can safely evolve and thus deal with ever-changing health knowledge using a two-level approach.

C. Similarity Analyser for Semantic Interoperability

Data interoperability goal is achieved when heterogeneous systems problems are resolved through ontology matching and through accurate mapping file generation and it helps in clinical message conversion from one standard to another. Healthcare standards play an important role in achieving interoperability between EHR systems.

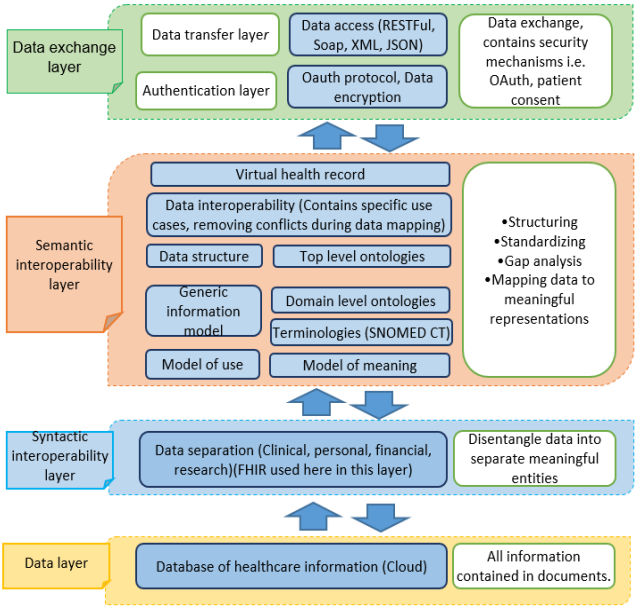

Fig. 1. Proposed Interoperability Framework.

Each healthcare system has its own goals and objectives. These include:

- HL7: Related to messaging.
- SNOMED CT: Related to terminologies.
- Open EHR and HL7 CDA: Clinical information and patient records.
- DICOM: Digital imaging and communication in medicine that is related to imaging and communication in medicine.

Two organizations are interoperable, if they are complaint with the same standards. Problem arises when different healthcare system uses different standards e.g. Open EHR complaint system cannot directly communicate with HL7 complaint system.

For this problem one solution is ontology mapping which is the process of eliminating the terminological and conceptual conflicts and discovering similarities and for this purpose similarity analyzer is introduced in our proposed framework and AI mapping techniques are used in similarity analyzer. So by using AI mapping, we can standardize clinical data records quickly and efficiently.

1. Working of Similarity Analyser Using AI:

Similarity analyzer performs various functions such as data structuring, data mapping, data modeling and conflict removal. Data is structured into various archetypes which provide specific information about an object such as blood pressure. Different types of conflicts are removed from the data to model data into common types which can be interpreted by different stakeholders in healthcare. So the main task of similarity analyser is that it takes the query from one hospital, analyse the standard or variation and then convert it into a standard format and reply back the required information in the desired standard.

For this purpose, the EHR data is classify into following types.

- Numeric Data.
- Textual Data.
- Images.

Numeric Data: For numeric data variations, we use Rule Based technique to convert the numeric data from one format to another. A simple example is that one hospital can use the patient's date of birth format like D/M/Y and the other hospital use the format like M/D/Y, so for this problem, Rule Based technique is used which work according to the query and convert the numeric data into desired format.

Textual Data: For Textual data, we classify data into two main components. One is unstructured data like physical examination reports, clinical laboratory reports, doctor's notes, summaries and other one is medical terminologies.

For unstructured data, Natural Language Processing (NLP) technique is used. NLP extract information from unstructured data and converts it into supplement and enriched structured medical data. NLP technique target at unstructured textual data and convert it into machine readable structured data by using Machine Learning (ML) techniques.

An NLP pipeline comprises of two main components. (1) Text processing and (2) classification. Through text processing, the NLP identifies the series of disease relevant keywords in the clinical notes, clinical laboratory notes based on patient's history database and then further analysis can be done on the reports and then these relevant keywords then enter and enrich the structured data and help in clinical decision making.

For Medical Terminologies, proposed similarity analyzer will use the Word2Vec AI technique. Word2Vec technique embed the words. Machine learning and deep learning cannot access text directly, which it need some sort of numeric representation so that the algorithm can process the data. In simple machine learning techniques, relationship between words cannot be reserved, so Word2Vec technique is used to embed the words.

Word2Vec is used to generate word embedding in a given text corpus. Word embedding means mapping of word in a vector space. So it preserves the relationship between words and deals with addition of new words in a vocabulary. The main objective is to cause the words that occur in similar context to have similar embedding.

Two algorithms CBOW and Skip gram are used to generate vectors from words. CBOW predicts the target words from context and Skip gram algorithm is used to predict the context words from target. So to improve the accuracy, we have to increase the training datasets, vector dimensions and window size but the drawback is that it increases the time duration.

Images: For images processing, our proposed similarity analyzer uses auto encoder technique which is a deep learning technique in which we add the images of different disease and then if there is a query arrived then it can predict the similarity in an unsupervised manner.

Flow of the proposed similarity analyzer is shown in the Fig. 2. The disease dataset is available to disease-data-server, from where disease-fetching API receive the data and then added to the disease-added-dataset. For the purpose of detection of similar words related to the given disease, proposed similarity analyzer part (Disease–Detection NLP) uses the data for the detection of the similarity or synonyms of the disease and full the query as a new report by giving the output as disease synonyms or similar words related to that disease given in the given text as input. Similarly, for image conversion (encoder-decoder AI technique) is used by the similarity analyzer to answer the query of the hospital/laboratory or any related authenticated person.

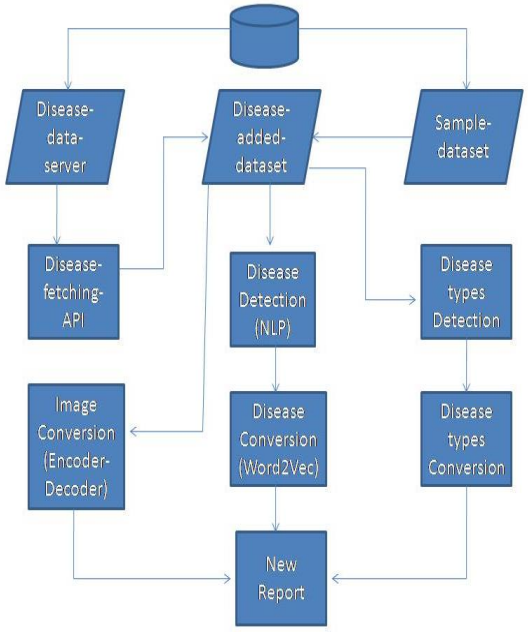

Fig. 2. Flow chart for Proposed Similarity Analyser

2. How Word2Vec Works

This section will describe the use of word2vec as an AI technique in our interoperability framework. As shown in Fig 2, Word2vec is part of the flowchart of the proposed Similarity Analyzer. Fig. 3 shows how Word2Vec works in our Similarity analyzer. On the input side, a word related to disease or a name of the disease is given from the disease data set in the form of text as an input and Word2vec embed the word as machine learning cannot access the word directly so there is a need of some numeric representation so that it can process the data.

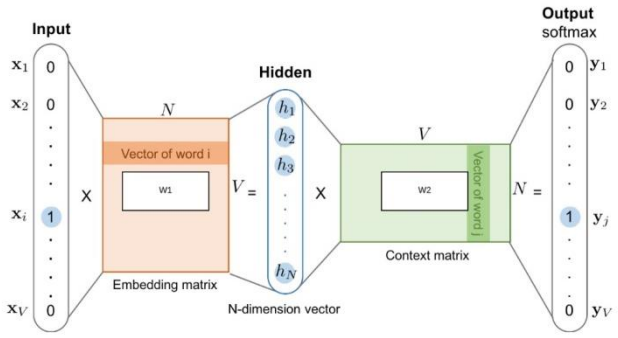

Fig. 3. Working of Word2Vec

The produce of a vector space in several hundred dimensions with each unique word in the text being assigned to the corresponding vector in the space so that the words that can share common context are located closely to each other in that vector space. So words having common context are located close to each other in the space and then as an output, similar words related to the disease word given as input are given as an output which is in the form of text or words.

3. How Doc2Vec Works

In this section, we will explain the working of doc2vec as an AI technique for similarity analysis in our interoperability framework. The Doc2Wec is an unsupervised learning algorithm which is used to develop representation of a document in a numerical form. As oppose to Word2Vec, the length of the document does not matter in Doc2Vec algorithm. However, the concept of Doc2Vec algorithm is heavily dependent of Word2Vec algorithm. The Doc2Vec algorithm introduces another vector in Word2Vec algorithm which is known as Paragraph ID (D) along with the word vector (W). The vector D is a unique reference to the document in the algorithm.

III. Results and Discussions

The analysis of two algorithms Word2Vec and Doc2Vec which are implemented in the proposed similarity analyzer framework. The analysis of the algorithms is made in terms of their accuracy of the semantic similarity of diseases and the processing time taken by the algorithms. Table1 provides results of Word2Vec algorithm when applied on a disease dataset. The algorithm predicts these words as semantically similar with the chosen disease "Pneumonia". The highest accuracy provided by the algorithm is 0.92 for the disease "Decreased-translucency".

The translucent lesion on the chest of a child observed in radiography due to fever or septic appearance causes the symptoms of the disease "Pneumonia". Fig.4 provides a visual representation of the semantic similarity of the diseases in the form of a scatter plot. The most similar disease appears closer to each other in the Fig 4.

Table 1: Word2Vec words similarity

Word	Accuracy
decreased-translucency	0.92
cough	0.83
infection	0.81
upper-respiratory-infection	0.75
bronchitis	0.74
lung-nodule	0.73

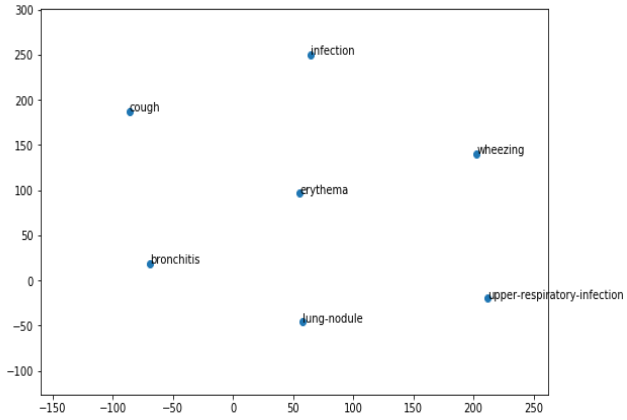

Fig. 4. Word2Vec scatter plot of diseases

Furthermore, when the same disease dataset is used with the Doc2Vec algorithm to find semantic similarity, the following Table2 illustrates the obtained disease accuracies. Doc2Vec also predicts the semantic similarity of disease "Pneumonia" with the "Translucency" disease. The semantic similarity of the disease from Table2 is visualized in Fig 5.

Table 2: Doc2Vec words similarity

Word	Accuracy
translucency	0.84
clonus	0.81
sputum	0.80
cachexia	0.74
infection	0.66
respiratory	0.57

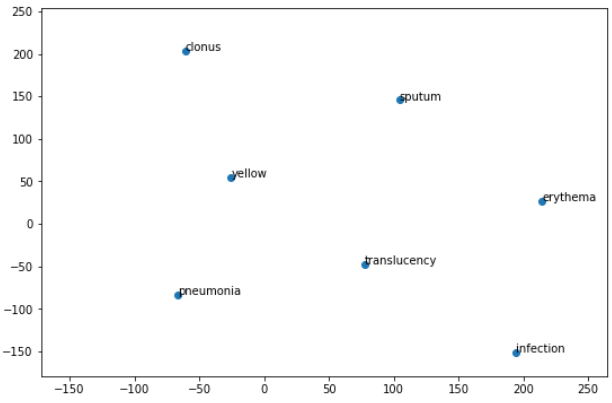

Fig. 5. Doc2vec scatter plot of diseases

Additionally, we also assess the processing time taken by the two algorithms i.e. Word2Vec and Doc2Vec. The parameter configurations and the processing time of each algorithm is indicated the following Table3. The algorithms Word2Vec and Doc2Vec are trained using the similar configuration for vector size and the number of epochs as illustrated in Table3. It is observed during the experimentation that the Doc2Vec algorithm requires more computation resources as compared to the algorithm Word2Vec.

The processing time taken by Word2Vec algorithm is 3 seconds while for Doc2Vec algorithm it is 16 seconds. This is due to the fact that the Doc2Vec algorithm finds the semantic similarity of the disease based on two vectors which are Paragraph ID represented as D and Word vector indicated as W.

Table 3: Word2Vec and Doc2Vec parameter configurations and processing time

Algorithm	Vector size	Epochs	Time(sec)
Doc2Vec	150	1000	16
Word2Vec	150	1000	3

IV. CONCLUSIONS

The proposition of electronic health records in the healthcare organization has opened up possibilities to migrate the patient records form the conventional, resource consuming paper based to the paperless electronic paradigm. However, the healthcare organizations face challenges to exchange patient health related information, laboratory reports with each other due to the existence of many standards. In this work, we propose a similarity analyzer framework to overcome the issue of semantic interoperability during the exchange of the electronic health records between organizations using AI techniques. We propose that the semantic interoperability is required in terms of numerical, textual and images based information. We provide detailed assessment of two algorithms Word2Vec and Doc2Vec to find the semantic similarity of the numerical and textual based disease dataset and show their accuracy and the resource consumption. The future work includes the implementing of the semantic interoperability in our proposed framework based on the images in the electronic health records.

REFERENCES

[1] A. Naveed, T. Sigwele, Y. F. Hu, M. Kamala, and J. Hou. (2018). Addressing Semantic Interoperability, Privacy and Security Concerns in Electronic Health Records. 2nd Annual Innovative Engineering Research Conferenve(AIREC) 2018. pp. 1-7.

[2] N. H. S. C. F. Health, "Snomed Ct," vol. 2012, no. 23/1/2012, 2012.

[3] D. H. Interoperability, "Digital Healthcare Interoperability," Vol. October, 2016.

[4] T. Sigwele, Y. F. Hu, M. Ali, J. Hou, M. Susanto, and H. Fitriawan. (2018). An intelligent edge computing based semantic gateway for healthcare systems interoperability and collaboration. 2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud), pp. 370-376

[5] C. Martinez-Costa, M. C. Legaz-Garcia, S. Schulz, and J. T. Fernandez-Breis. (2014). Ontology-based infrastructure for a meaningful EHR representation and use," 2014 IEEE-EMBS Int. Conf. Biomed. Heal. Informatics, BHI 2014. pp. 535–538.

[6] C. N. Mead. (2006). Data interchange standards in healthcare IT--computable semantic interoperability: now possible but still difficult, do we really need a better mousetrap? J. Healthc. Inf. Manag., vol. 20, no. 1, pp. 71–78.

[7] S. Bhartiya, and D. Mehrotra. (2016). Challenges to Sharing of Electronic Health Records in Interoperable Environments. HIMMS Asia Pacific. Vol. July, pp.1–2.

[8] S. Bhartiya, D. Mehrotra, and A. Girdhar. (2016). Issues in Achieving Complete Interoperability while Sharing Electronic Health Records. Procedia Comput. Sci., vol. 78, pp. 192–198.

[9] O. Iroju, A. Soriyan, I. Gambo, and J. Olaleke. (2013). Interoperability in Healthcare: Benefits, Challenges and Resolutions. Int. J. Innov. Appl. Stud., vol. 3, no. 1, pp. 262–270.

[10] Z. Bouanani-Oukhaled *et al.* (2017). Ontological Model for EHR interoperability. [Online] HAL Id: hal-01457845. pp. 1-6. Available: https://hal.archives-ouvertes.fr/hal-01457845.

[11] S. S. El-Atawy and M. E. Khalefa. (2016). Building an Ontology-Based Electronic Health Record System. Proc. 2nd Africa Middle East Conf. Softw. Eng. - AMECSE'16. Vol. May 2016, pp. 40–45.

[12] W. A. Khan et al. (2012). Achieving Interoperability among Healthcare Standards : Building Semantic Mappings at Models Level. Proceedings ICUIMC '12: The 6th Int. Conference on Ubiquitous Information Management and Communication. pp. 1-9. Available: https://dl.acm.org/doi/pdf/10.1145/2184751.2184868

[13] M. Jaulent, D. Leprovost, J. Charlet, and R. Choquet. (2018). Semantic interoperability challenges to process large amount of data perspectives in forensic and legal medicine. J. Forensic Leg. Med., vol. 57, pp. 19–23.