eCREAM
enabling Clinical Research
in Emergency and Acute care Medicine

STUDY PROTOCOL

NLP-DeVal: Development and validation of a natural language processing tool to enable clinical research in emergency and acute care medicine: retrospective cohort study

Version 1.2

Ranica, February 23rd, 2024

clinicaltrials.gov: NCT06240572

.

eCREAM Coordinating Center
Istituto di Ricerche Farmacologiche Mario Negri IRCCS
Villa Camozzi. via GB Camozzi 3 - 24020 Ranica (Bergamo)

tel. +39 035 4535389 - fax +39 035 4535354 - e-mail: ecream@marionegri.it – Website: ecreamproject.eu

SUMMARY

SYNOPSIS **4**

1. INTRODUCTION **5**

1.1. BACKGROUND AND RATIONALE 5

1.2. THE ECREAM PROJECT 5

2. OBJECTIVE OF THE STUDY **6**

3. STUDY DESIGN **6**

4. PARTICIPATING CENTERS AND STUDY POPULATION **6**

4.1. PARTICIPATING CENTERS 6

4.2. STUDY POPULATION 6

5. STUDY PLAN AND METHODS **6**

5.1. DEVELOPING AND VALIDATING THE ECREAM_LM 6

5.2. STUDY DURATION 7

5.3. DATA COLLECTION AND ANONYMIZATION 7

5.4. STATISTICAL ANALYSIS 9

5.5. SAMPLE SIZE 10

5.6. DATA PROCESSING 10

6. ADVERSE EVENTS **11**

7. STUDY MONITORING **11**

8. ETHICAL CONSIDERATIONS **11**

8.1. THE LEGAL BASIS FOR DATA PROCESSING 11

8.2. INSURANCE 12

8.3. ETHICS COMMITTEE 12

8.4. PUBLICATION OF DATA 12

8.5. DATA SHARING 12

9. ROLES OF RESPONSIBILITY **12**

10. STUDY SUPPORT **13**

11. REFERENCES **13**

STUDY CONTACTS

Principal Investigator	Guido Bertolini - Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Italy
Responsible for Coordination of Ethics Committees	Mihaela Matei - ECRIN (European Clinical Research Infrastructure Network)
Scientific Committee	Guido Bertolini, Rita Banzi - Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Italy Jacek Gorka - Uniwersytet Jagiellonski, Poland Alberto Lavelli, Bernardo Magnini – Fondazione Bruno Kessler, Italy Federica Moretti – Orobix Life S.R.L., Italy George Notas - Dioikhsh Ygeionomikhs perifereias Krhths, Greece Peter Mitro - Univerzita Pavla Jozefa Safarika v Kosiciach, Slovakia Gregor Prosen - Univerzitetni klinicni center Maribor, Slovenia Pankaj Sharma - Ashford and St Peter's hospitals NHS foundation trust, United Kingdom
Technical partners	Astir S.R.L., Italy Fondazione Bruno Kessler (FBK), Italy Orobix Life S.R.L., Italy
Project Management	Giulia Irene Ghilardi, Chiara Pandolfini, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Italy Tel +39 035 4535389/351, +39 02 39014245 giulia_irene.ghilardi@marionegri.it, chiara.pandolfini@marionegri.it
Data management and Statistical analysis	Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Italy Fondazione Bruno Kessler, Italy Orobix Life S.R.L., Italy
IT support	Astir S.R.L., Italy
National study contacts	IT: Giulia Irene Ghilardi, giulia_irene.ghilardi@marionegri.it Chiara Pandolfini, chiara.pandolfini@marionegri.it Istituto di Ricerche Farmacologiche "Mario Negri IRCCS" PL: Jacek Gorka, gorkajacek1@gmail.com Uniwersytet Jagiellonski SI: Maja Molan, maja.molan@ukc-mb.si Univerzitetni klinicni center Maribor SK: Nikola Jurekova, nikola.jurekova@upjs.sk Univerzita Pavla Jozefa Safarika v Kosiciach UK: Isaac John, isaac.john@nhs.net Ashford and St Peter's hospitals NHS foundation trust EL: Aimilia Magkanaraki, amagkanaraki@hc-crete.gr Dioikhsh Ygeionomikhs perifereias Krhths

SYNOPSIS

BACKGROUND AND RATIONALE	Doing clinical research in emergency medicine is as difficult as important. It is difficult because the large number of patients to be treated and the chronic shortage of staff make any data collection impractical. It is important because it allows emergency physicians and nurses to base their practice on evidence produced in the very distinct context of care in which they work and not, as is predominantly the case today, in contexts far removed from it. The only way to fill the gap between the need for research and the availability of data is to extract the latter directly from the electronic health records (EHR) of emergency departments (EDs). This is a difficult task, however, because a large part of the useful information is contained in an unstructured format, in free text notes. Today, large-scale language models that can accurately interpret natural language are available. These models are trained on huge amounts of general data, mostly taken from the Internet, so their performance in more specialized areas, such as the medical domain, may not be optimal.
OBJECTIVES OF THE STUDY	To develop and validate a language model (called *eCREAM_LM*) for the six languages of the project that is able to interpret EHR contents and extract relevant information for research purposes.
STUDY DESCRIPTION	33 centers from 7 different countries (Italy, Poland, Greece, Slovakia, Slovenia UK, and Switzerland) will participate. The creation of the *eCREAM_LM* language model will entail training and fine-tuning different language models through exposure to both a massive amount (billions) of medical texts derived from the literature and a large quantity of non-annotated (millions) and annotated (5,000) clinical notes derived from EHRs. The validation of *eCREAM_LM* will be done by assessing its concordance with a panel of experts in extracting crucial information from 3,000 clinical notes. The study will last 24 months and will involve retrospective data collection.
STATISTICAL ANALYSIS	In validating *eCREAM_LM*, the agreement between experts and *eCREAM_LM* in completing eCREAM's virtual case report form (vCRF) in a sample of 1,000 notes for each study language will be evaluated. Concordance will be assessed for each variable in the vCRF using Cohen's κ as a measure of agreement. *eCREAM_LM* will be considered valid if Cohen's κ is greater than 0.75.
ETHICAL CONSIDERATIONS	All the information collected will be anonymized and will therefore not require a specific legal basis to be treated.
FUNDING	The study is supported by the European Commission (Project number: 101057726), UKRI (UK Research and Innovation), and SERI (Swiss State Secretariat for Education, Research and Innovation).

1. INTRODUCTION

1.1. Background and rationale

Conducting clinical and quality-of-care assessment research in emergency medicine is as difficult as important. It is difficult because the vast number of patients who visit an emergency department (ED), 60-90,000 patients per year in an average-sized center, and the staff shortages that chronically plague these departments make ad hoc data collection impractical. It is important because, in the end, research enables emergency physicians and nurses to base their practice on evidence obtained in their own, unique setting, as opposed to evidence obtained in far-removed contexts, as is commonly the case today.

The only way to fill the gap between the need for clinical research and the availability of robust data is to directly extract such data from the ED's electronic health records (EHRs), avoiding dedicated, time-consuming data collection. Nonetheless, obtaining consistent data from EHRs is a complex task. While part of the data registered in EHRs is structured (e.g., lab test results and vital parameters) and therefore easy to retrieve, the most useful patient information is often in free text form (e.g., presence of signs and symptoms, suspected and confirmed diagnosis, anamnesis). Such circumstances and needs require a reliable natural language processing (NLP) tool to derive highly consistent data from free text.

Large language models (LLMs) able to thoroughly interpret natural language are now available. These models have achieved remarkable performance on a wide range of language-related tasks and are capable of accurately extracting relevant information even from conversational texts. There are, however, significant limitations in using these models for the purpose we have just indicated. First, their general performance values, and therefore their overall reliability, are not known. Second, these models are trained on massive datasets from the Internet, which means they have been exposed to a broad range of general knowledge. While this enables them to provide fairly accurate performance on common topics, they tend to falter in more specialized areas, such as the one relevant to this study. Third, the largest and best-performing models are proprietary, and their integration with other software is expensive. Thus, using best-performing LLSs is economically and computationally unsustainable in the context of continuing multicenter research projects that potentially involve a large number of patients and centers.

The present study (NLP-DeVal), which refers to the first pillar of the eCREAM project, aims to train and validate an LLM able to extract accurate information from the EHRs of emergency departments.

1.2. The eCREAM project

eCREAM is a European project coordinated by the Laboratory of Clinical Epidemiology at the Istituto di Ricerche Farmacologiche Mario Negri IRCCS (IRFMN). It involves 11 partners in 8 countries (France, Greece, Italy, Poland, Slovakia, Slovenia, Switzerland, and the United Kingdom).

eCREAM is structured in 5 pillars: 1) develop a natural language processing tool tailored to the typical notes made by emergency physicians and nurses; 2) develop an IT system able to capture data contained in the EHRs in use in different EDs and other administrative databases; 3) develop a new electronic health record for EDs, designed to meet organization, clinical practice, and clinical research needs; 4) take on two different use cases to test and validate the whole system in the field: the assessment of ED propensity to hospitalize patients (use case 1), and the development of a dashboard to be used by citizens, policymakers, and healthcare providers to improve the quality of care in the ED (use case 2); 5) FAIRify (i.e., make data Findable, Accessible, Interoperable, and Re-usable) the

established databases for clinicians, researchers, and health policymakers and citizens, respecting the European and national legislations.

eCREAM has been financed by the European Commission under the *Horizon Europe* program (contract no. 101057726), UKRI (UK Research and Innovation), and SERI (Swiss State Secretariat for Education, Research and Innovation).

2. OBJECTIVE OF THE STUDY

The objective of this study is to develop and validate a state-of-the-art language model (called *eCREAM_LM*) for the six languages of the project (English, Greek, Italian, Polish, Slovakian, and Slovenian), able to interpret EHR contents and extract crucial information from them, which can be used to make accurate analyses and predictions.

3. STUDY DESIGN

Observational, multicenter, retrospective study.

4. PARTICIPATING CENTERS AND STUDY POPULATION

4.1. Participating centers

The hospitals adhering to the eCREAM project will participate in the study. Overall, 33 centers will be included: 16 from Italy, 4 from Poland, 3 from Greece, Slovakia, Slovenia and the UK, and 1 from Switzerland. The complete list of participating centers is reported in Annex 1. The centers will not receive any incentive for participating in the study, but expenses incurred for participation will be covered by project funds.

4.2. Study population

Adult patients who arrived at the participating hospitals between January 1, 2021, and December 31, 2023, will be eligible for the study.

5. STUDY PLAN AND METHODS

5.1. Developing and validating the eCREAM_LM

The *eCREAM_LM* will be a language model for the six languages of the project (English, Greek, Italian, Polish, Slovakian, and Slovenian), able to interpret the content of EHRs and, therefore, to extract information necessary for research purposes. It will be an NLP model based on deep learning architectures, developed by training or fine-tuning the best LLM among those available open source, or through a different solution that has become available in the meantime. Indeed, the world of LLMs is constantly and rapidly evolving, and eCREAM's goal is to adopt a solution that guarantees the highest possible level of reliability.

Since the process required to train and evaluate language models is lengthy and complex, the development of the *eCREAM_LM* will proceed in partially parallel steps. First, the LLM candidates will be trained by exposing them to a massive amount (billions) of medical texts derived from scientific literature or other publicly available sources, such as the Internet. This phase will continue for the duration of the project, during which we will gradually accumulate as many texts as possible. Concurrently, each candidate model will be exposed to a large amount (millions) of non-annotated clinical notes obtained from the EHRs in use in the participating hospitals. The final step will consist of

fine-tuning the LLM candidates using a large quantity of clinical notes (at least 5,000) obtained from the EHRs of the participating hospitals and annotated by trained experts. Specifically, the annotation procedure refers to the process by which experts will read the anonymized notes and will fill in those items of the eCREAM virtual case report form (vCRF) that are referred to in the notes, extrapolating the necessary data, so that the *eCREAM_LM* can learn to do the same job. The eCREAM vCRF is the collection of the variables considered important for predicting the hospitalization of patients with dyspnea or transient loss of consciousness (TLOC), according to the eCREAM Use Case 1 study protocol. These variables were identified through a Delphi-modified method and a literature review by an international panel of experts. Since no active data collection is requested from the participating centers, as all necessary information is obtained directly and automatically from the EHRs in use in the EDs by the *eCREAM_LM*, we refer to this list of variables as the eCREAM virtual case report form; it is reported in Annex 2.

After the final step of development, the *eCREAM_LM* will be validated. A set of 3,000 additional clinical notes will be annotated in the same way described for the fine-tuning process in the development phase. These notes will be submitted to the *eCREAM_LM* with a request to complete the eCREAM vCRF. The concordance in completing the vCRF between *eCREAM_LM* and the trained emergency physicians who annotated the clinical notes will represent the measure of the final validation of *eCREAM_LM*.

5.2. Study duration

The study will last 24 months: the first six will be used to install the anonymization software in the various participating hospitals and to retrieve the clinical notes from the EHRs; the following six months will be used to train the candidate models with non-annotated clinical notes, and to annotate the clinical notes that will be used for the final phase of *eCREAM_LM* development and its validation. The remaining 12 months will be devoted to fine-tuning the *eCREAM_LM*, its validation, data analysis, research reports production, and results dissemination.

5.3. Data collection and anonymization

A large sample (millions) of free texts extracted from EHRs is needed to develop the *eCREAM_LM*. Each participating hospital will provide the clinical notes contained in the EHRs of approximately 90,000 to 300,000 patients, depending on the size of the center, who visited the hospital between 2021 and 2023. The notes will be stripped of any reference to the patient (first name, last name, date of birth, address, phone number, etc.) and to the context (hospital, date and time of arrival at the center, date and time of visit, hospitalization decision, date and time of discharge/hospitalization, etc.). Furthermore, the notes referring to different aspects of the same patient (e.g., medical history, clinical examination, test results) will be separated from each other so that it will be impossible to reconstruct the entire patient profile. This process minimizes the likelihood of identifying patients and maximizes the protection of patients' rights. The possibility of identifying a patient within a database depends on the unicity of his or her characteristics with respect to the characteristics of the other subjects contained in the same dataset. Accordingly, the probability of having unique patients in a dataset increases with the amount of information available and decreases with the size of the database. Removing all personal and contextual information from the clinical notes and separating each note from the others related to the same patient results in only a few patient characteristics being reported in each note. Moreover, the data gathered from different hospitals in the same country will be merged in order to have one large database per language. This process effectively resets the probability that there are individuals uniquely identifiable from the notes. At the same time, this also ensures maximum protection of the

patients' right not to have their personal characteristics revealed. In fact, to identify a specific patient in a database, one should already know all the characteristics that make that patient unique in that database. Since the process we have described results in individual notes containing only a few pieces of information about the patient, these notes would likely not contain any additional information beyond what is needed to identify the patient as unique. This means that if someone were able to infer that a particular note refers to a known person (an already virtually impossible eventuality), he or she would not find any information in that note that is not already known. Furthermore, state-of-the-art methods will protect the database, making a data breach highly unlikely.

While the process of separating the notes referring to different aspects of the same patient reduces, if not nullifies, the possibility of re-identifying patients, the process increases the quantity of notes that need to be annotated for the fine-tuning of the candidate language models. Separating the notes into those referring only to a single aspect of the health record (i.e., medical history, clinical examination, and clinical course) reduces the presence of the information needed by the annotator to fill in all the vCRF's data items. For example, information on whether the patient has diabetes (one of the vCRF items) can be found in the medical history notes and not in the clinical examination ones. Similarly, information on whether the patient was confused (another item of the vCRF) can be found in the clinical examination notes, but probably not in the clinical course notes. Therefore, in order to assess a sufficient number of notes to cover all the items in the vCRF, over 1,500 notes will have to be annotated for each of the three groups, for a total of 5,000 notes.

The procedures described above, i.e., stripping the notes provided by the hospitals of any reference to the individual patients and contexts, and separating the notes referring to different aspects of the same patient, ensure that the probability of patient identification is less than negligible. We cannot rule out, however, the possibility that the notes contain information about third parties. The physician or nurse, for example, may write down the names of patient relatives with their phone numbers or address in case they need to be contacted. Two participating Italian centers reviewed their clinical notes and estimated a 10% occurrence of such references to third parties.

To remove the possible presence of information about third parties from the database, we will install, in each participating hospital, a certified anonymization software called AnonymAI. This software is specifically designed to identify and remove data entities referring to specific types of personal data from free text. We will use AnonymAI to remove the following: first name, last name, phone number, address, location, and parental relationship.

Annex 3 contains the description of the AnonymAI software provided by its manufacturer, with the software's level of performance in relation to our purpose. We reproduce here the most informative table containing the probability of failing to anonymize a subject mentioned in a note, as well as the number of notes containing a reference to a subject that need to be analyzed to have one subject not anonymized.

Combination of identifiers	Probability of not anonymizing a subject in a note	Number of notes to analyze to have one non-anonymized subject
Full name (first + last names)	0.00045	2,222
Full name + phone number	0.000009	111,111
Full name + full address	1.98 E-7	5,050,505
Full name + phone number + full address	3.96 E-9	252,525,253
Full name + parental relationship	1.85 E-5	54,054
Full name + phone number + parental relationship	3.69 E-7	2,710,027
Full name + full address + parental relationship	8.118 E-9	123,183,050
Full name + phone number + full address + parental relationship	1.6 E-10	6,250,000,000

Of note, knowing that a person with a given first and last name was cited in the clinical notes of a hospital in three years does not reveal anything in itself, i.e., it does not violate anyone's privacy. Consequently, we need to concentrate on the probability of not anonymizing the combination of, for example, full name and phone number in a free text note. Indeed, this means possibly knowing the phone number of a person whose first and last name we also know. We must use the word "possibly" because we cannot be sure that the phone number not removed from the text belongs to the person whose full name was not removed either. The table above shows that such an event occurs once in approximately 111,000 instances. Considering that, in each country, we will retrieve 400,000-750,000 notes and that only 10% of them will contain a reference to a third party (which may not necessarily involve the first name, last name, and phone number), we can state that, on average, such an event will not occur in any of the participating countries. Furthermore, the riskiest situation in terms of subject-rights protection would be if a hacker were able to identify a subject through a given combination of identifiers (e.g., full name and phone number) and thus learn that he or she had a relative admitted to a hospital with a defined clinical scenario. Such an event could occur at most once in approximately 2,700,000 cases. We can therefore affirm that the occurrence of such an event would be virtually impossible since, for each language, we will collect a maximum of 75,000 notes containing a reference to a third party (10% of 750,000 notes), which is very far from the 2,700,000 notes needed to result in just one such event.

Once anonymized, data will be centralized to the eCREAM-NLP server at the IRFMN in Milan, Italy, for subsequent analyses.

5.4. Statistical analysis

In the *eCREAM_LM* validation, we will assess the concordance between expert emergency physicians and the *eCREAM_LM* itself in filling in the vCRF developed for the eCREAM Use Case 1 study protocol (see Annex 2). The

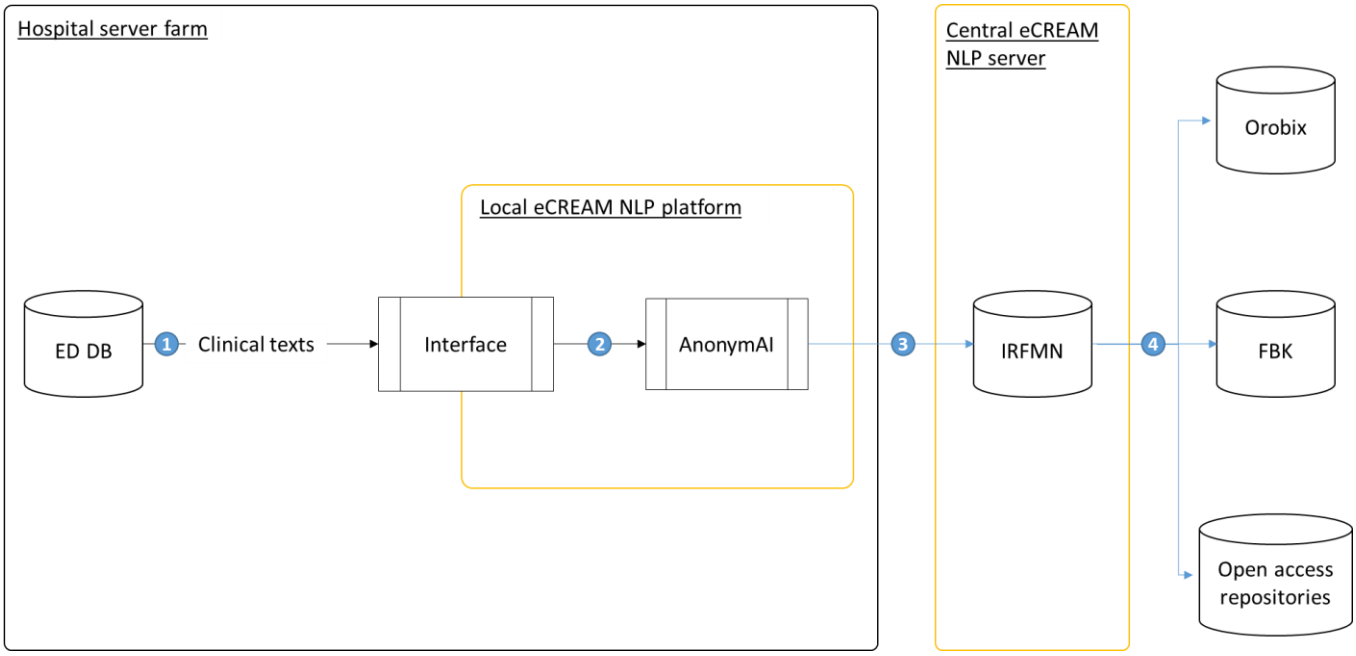
data will refer to a sample of 3,000 notes for each study language. Concordance will be assessed for each variable of the vCRF using Cohen's κ as a measure of agreement. The *eCREAM_LM* will be considered valid if Cohen's κ is greater than 0.75.

5.5. Sample size

The sample size is computed for the validation phase. Assuming an excellent agreement (κ=0.80) between eCREAM_LM and the experienced emergency physicians in completing the vCRF, a sample of at least 735 notes will be necessary to achieve sufficient precision to guarantee a good agreement (lower confidence limit of 95% confidence interval of Cohen's κ greater than 0.75). This number is the maximum sample size obtained under different scenarios involving a different number of categories (2 to 5) for each variable and different marginal distributions of the categories in the sample, including balanced distributions (e.g., 5 categories with 20% of the sample in each category) and very imbalanced results (e.g., 5 categories with 1.8%, 7.3%, 16.4%, 29.1% and 45.5% of the sample). Since information of interest may be missing in some notes, we will perform the data validation assessment on 1,000 notes. For the same reason explained in section 5.3 on fine-tuning models (i.e., to cover all the items included in the vCRF), it will be necessary to multiply this number by 3, annotating a total of 3,000 notes for the validation phase.

5.6. Data processing

The figure below represents the data processing that will be carried out in the context of this project.



The first data processing step will take place within the participating hospitals' information systems (*Hospital server farm* in the figure), where the *Local eCREAM NLP platform* specially developed for the project will be installed. This step consists of extracting all the free text notes from the emergency department's EHR after removing all references to the patient and context, as described in paragraph 5.3, and transmitting them to the *Local eCREAM NLP platform* through an ad hoc *Interface*. Integration APIs between hospital systems and the *Local eCREAM NLP*

platform will be exposed on the intranet, not the Internet. Thus, the transmission of data that has not yet been processed by the eCREAM platform will take place within the hospital's internal network.

The *Local eCREAM NLP platform* is designed to remove all possible references to third parties (like patients' relatives) from these notes. This process, achieved through the *AnonymAI* software integrated into the platform (second step), results in anonymized clinical notes. Databases and file systems that store these notes within the *Local eCREAM NLP platform* are encrypted. The clinical notes are retained in such archives as long as necessary for further processing and subsequent transmission (third step) to the *Central eCREAM NLP server*, located on the premises of the Istituto di Ricerche Farmacologiche Mario Negri IRCCS (IRFMN) in Milan, Italy, the eCREAM coordinator. This transmission will be carried out through secure protocols. At the end of the transmission process, the notes will be physically deleted from local repositories.

Finally, the fourth step will consist of transmitting the anonymized free texts to the Fondazione Bruno Kessler (FBK) in Trento and Orobix Life (Orobix) in Bergamo, the eCREAM partners that will be in charge of *eCREAM_LM* development, and to the main European platforms for language resource sharing, including the European Language Grid repository (https://live.europeanlanguage-grid.eu), the CLARIN Virtual Language Observatory (https://vlo.clarin.eu/?2), and the European Language Equality (ELE) initiative (https://european-language-equality.eu). For more details on data processing and protection, see Annex 4.

6. ADVERSE EVENTS

As an observational retrospective study, participants will be treated according to routine clinical practice. No specific adverse event data collection will therefore be set up.

7. STUDY MONITORING

While the coordinating center will oversee data collection centrally, in-site study monitoring will not be performed.

8. ETHICAL CONSIDERATIONS

8.1. The legal basis for data processing

All collected information will be anonymized according to the process described in paragraph 5.3. Anonymization is prescribed, whenever possible, by EU Regulation 2016/679 when data are used for scientific research purposes. More specifically, Article 9, paragraph 2, letter j provides that the data subject's consent is not required when the processing is necessary, among other conditions, for scientific research purposes, in accordance with Article 89(1), based on Union or national law, provided that it is proportionate to the purpose pursued, respects the essence of the right to data protection, and provides for appropriate and specific measures to protect the fundamental rights and interests of the data subject. Paragraph 1 of Article 89, chapter IX, which is recalled here, states that "Processing for archiving purposes in the public interest, scientific or historical research or statistical purposes shall be subject to appropriate safeguards for the rights and freedoms of the data subject, in accordance with this Regulation. Such safeguards shall ensure that technical and organizational measures are in place, in particular to ensure compliance with the principle of data minimization. Such measures may include pseudonymization, provided that the purposes in question can be achieved in this manner. Where they can be achieved by further processing that does not allow, or no longer allows, the data subject to be identified, those purposes must be achieved in this manner."

As controllers of their own data, the individual centers will be responsible for anonymization according to the process described in paragraph 5.3, with the support of Astir (a partner of the eCREAM consortium) if necessary. Once anonymized, data are not subject to the European General Data Protection Regulation (GDPR) 2016/679 and thus do not require a specific legal basis to be treated.

8.2. Insurance

Considering the retrospective design and the anonymity of data collected, no study-specific insurance coverage is required for this study.

8.3. Ethics Committee

According to the national legislation for observational studies, this study protocol will be submitted to all participating centers' ethics committees.

8.4. Publication of Data

Publications will be decided on by the Steering Committee (SC), which includes the principal investigator, the co-investigators, and the people responsible for the national coordinating centers. The authors to be reported on the front page will be selected based on their specific contributions to the project. All manuscripts will include an appropriate acknowledgment section, mentioning all investigators who contributed to the study and all supporting agencies, making explicit reference to the European grant that enabled the project to be carried out. Research results will be published irrespective of the findings.

8.5. Data Sharing

The anonymized data derived from the annotated and non-annotated clinical texts used to train the NLP models that will be collected in this study represent a significant linguistic dataset resource. They will consequently be transmitted to the main European platforms for language resource sharing.

9. **ROLES OF RESPONSIBILITY**

The Istituto di Ricerche Farmacologiche Mario Negri IRCCS (IRFMN) is responsible for overseeing the activities requested to ensure that the study will be conducted in compliance with the protocol, applicable regulatory requirements, and international guidelines. The PI and scientific committee will have scientific and operational responsibility.

FBK and Orobix are responsible for developing the *eCREAM_LM*.

Astir could act as a data processor for the individual centers if needed to ensure data processing, including anonymization, as described in paragraph 5.3.

The individual centers participating in the project are responsible for their own data and for applying the data processing, including anonymization, as described in paragraph 5.3.

Each country coordinator/national study contact will be responsible for obtaining ethical, regulatory, and legal authorizations in compliance with the European and local provisions, including the GDPR and national data protection regulations.

10. **STUDY SUPPORT**

The study is supported by the European Commission (Project number: 101057726), UKRI (UK Research and Innovation), and SERI (Swiss State Secretariat for Education, Research and Innovation).

11. **REFERENCES**

1. Smulowitz PB, O'Malley AJ, McWilliams JM, Zaborski L, Landon BE. Variation in Rates of Hospital Admission from the Emergency Department Among Medicare Patients at the Regional, Hospital, and Physician Levels. *Ann Emerg Med*. Published online June 18, 2021:S0196-0644(21)00233-X. doi:10.1016/j.annemergmed.2021.03.020

2. Smulowitz PB, O'Malley AJ, Zaborski L, McWilliams JM, Landon BE. Variation In Emergency Department Admission Rates Among Medicare Patients: Does The Physician Matter? *Health Affairs*. 2021;40(2):251-257. doi:10.1377/hlthaff.2020.00670

3. Gorski JK, Batt RJ, Otles E, Shah MN, Hamedani AG, Patterson BW. The Impact of Emergency Department Census on the Decision to Admit. *Acad Emerg Med*. 2017;24(1):13-21. doi:10.1111/acem.13103

4. Warner LSH, Galarraga JE, Litvak O, Davis S, Granovsky M, Pines JM. The Impact of Hospital and Patient Factors on the Emergency Department Decision to Admit. *J Emerg Med*. 2018;54(2):249-257.e1. doi:10.1016/j.jemermed.2017.11.024

5. Khojah I, Li S, Luo Q, et al. The relative contribution of provider and ED-level factors to variation among the top 15 reasons for ED admission. *Am J Emerg Med*. 2017;35(9):1291-1297. doi:10.1016/j.ajem.2017.03.074

6. Bonetti M, Melani C. Il ruolo degli accessi impropri in pronto soccorso nella provincia autonoma di Bolzano. *BEN*. Published online April 2019. Accessed March 17, 2023. https://www.epicentro.iss.it/ben/2019/aprile/accessi-impropri-ps-bolzano

7. McKenna P, Heslin SM, Viccellio P, Mallon WK, Hernandez C, Morley EJ. Emergency department and hospital crowding: causes, consequences, and cures. *Clin Exp Emerg Med*. 2019;6(3):189-195. doi:10.15441/ceem.18.022

8. Eurispes, Enpam. *Il termometro della salute. 1° Rapporto sul sistema sanitario*. 2018th ed. Minerva Edizioni Accessed March 17, 2023. https://www.ibs.it/termometro-della-salute-1-rapporto-libro-vari/e/9788833240046?lgw_code=1122-B9788833240046&gclid=Cj0KCQjwn9CgBhDjARIsAD15h0BXv3YYp1a-br1KLYfiMx_Mv3F3PE7F8oFHPgmUrGWMuDfcTuJ1TJUaAjm7EALw_wcB

9. Cliff BQ, Avanceña ALV, Hirth RA, Lee SYD. The Impact of Choosing Wisely Interventions on Low-Value Medical Services: A Systematic Review. *Milbank Q*. 2021;99(4):1024-1058. doi:10.1111/1468-0009.12531

10. Ciapponi A, Fernandez Nievas SE, Seijo M, et al. Reducing medication errors for adults in hospital settings. *Cochrane Database Syst Rev*. 2021;11(11):CD009985. doi:10.1002/14651858.CD009985.pub2

11. Shrank WH, Rogstad TL, Parekh N. Waste in the US Health Care System: Estimated Costs and Potential for Savings. *JAMA*. 2019;322(15):1501-1509. doi:10.1001/jama.2019.13978

12. Ogrinc G, Dolansky M, Berman AJ, Chambers DA, Davies L. Different approaches to making and testing change in healthcare. *BMJ*. 2021;374:n1010. doi:10.1136/bmj.n1010

13. Navathe AS, Volpp KG, Bond AM, et al. Assessing The Effectiveness Of Peer Comparisons As A Way To Improve Health Care Quality. *Health Affairs*. 2020;39(5):852-861. doi:10.1377/hlthaff.2019.01061

14. Winickoff RN, Coltin KL, Morgan MM, Buxbaum RC, Barnett GO. Improving Physician Performance through Peer Comparison Feedback. *Medical Care*. 1984;22(6):527-534.

15. Abualenain J, Frohna WJ, Shesser R, Ding R, Smith M, Pines JM. Emergency department physician-level and hospital-level variation in admission rates. *Ann Emerg Med*. 2013;61(6):638-643.

doi:10.1016/j.annemergmed.2013.01.016

16. Nattino G, Finazzi S, Bertolini G. A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes. *Statistics in medicine*. 2014;33(14):2390-2407. doi:10.1002/sim.6100

17. Nattino G, Finazzi S, Bertolini G. A new test and graphical tool to assess the goodness of fit of logistic regression models. *Statistics in Medicine*. 2016;35(5):709-720. doi:10.1002/sim.6744

18. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373-1379.

19. Pines JM, Mutter RL, Zocchi MS. Variation in Emergency Department Admission Rates Across the United States. *Med Care Res Rev*. 2013;70(2):218-231. doi:10.1177/1077558712470565

20. Kim SW, Li JY, Hakendorf P, Teubner DJ, Ben-Tovim DI, Thompson CH. Predicting admission of patients by their presentation to the emergency department. *Emerg Med Australas*. 2014;26(4):361-367. doi:10.1111/1742-6723.12252

21. Abir M, Goldstick JE, Malsberger R, et al. Evaluating the impact of emergency department crowding on disposition patterns and outcomes of discharged patients. *International Journal of Emergency Medicine*. 2019;12(1):4. doi:10.1186/s12245-019-0223-1

22. Brink A, Alsma J, Attekum LA van, et al. Predicting inhospital admission at the emergency department: a systematic review. *Emerg Med J*. 2022;39(3):191-198. doi:10.1136/emermed-2020-210902

23. Pines JM, Galarraga JE, Granvosky M, Litvak O, Davis S, Warner LH. How (ED) admission decisions differ when the same physician works in two different emergency department. *The American Journal of Emergency Medicine*. 2017;35(7):970-973. doi:10.1016/j.ajem.2017.01.067

24. Hosmer DW, Lemeshow S. Confidence interval estimates of an index of quality performance based on logistic regression models. *Stat Med*. 1995;14(19):2161-2172.

25. Solbiati M, Casazza G, Dipaola F, et al. Syncope recurrence and mortality: a systematic review. *EP Europace*. 2015;17(2):300-308. doi:10.1093/europace/euu327

26. Laribi S, Keijzers G, van Meer O, et al. Epidemiology of patients presenting with dyspnea to emergency departments in Europe and the Asia-Pacific region. *Eur J Emerg Med*. 2019;26(5):345-349. doi:10.1097/MEJ.0000000000000571

27. Foltran F, Baldi I, Bertolini G, Merletti F, Gregori D. Monitoring the performance of intensive care units using the variable life-adjusted display: a simulation study to explore its applicability and efficiency. *J Eval Clin Pract*. 2009;15(3):506-513. doi:10.1111/j.1365-2753.2008.01052.x

NLP-DeVal: Development and validation of a natural language processing tool to enable clinical research in emergency and acute care medicine: retrospective cohort study

ANNEX 1

Elenco centri partecipanti e relative responsabili

Version 1.1

Ranica, February 7, 2024

eCREAM Coordinating Center
Istituto di Ricerche Farmacologiche Mario Negri IRCCS
Villa Camozzi. via GB Camozzi 3 - 24020 Ranica (Bergamo)

tel. +39 035 4535389 - fax +39 035 4535354 - e-mail: ecream@marionegri.it – Website ecreamproject.eu

Ospedale San Giovanni Bosco, Torino (COORDINATORE)	Dott. Franco Aprà Direttore S.C. Medicina d'Emergenza e Urgenza 2 Ospedale S. Giovanni Bosco, Torino franco.apra@aslcittaditorino.it
Azienda Ospedaliera SS. Antonio e Biagio e C. Arrigo, Alessandria	Dott. Riccardo Boverio Direttore Medicina e Chirurgia d'Accettazione e Urgenza AO SS. Antonio e Biagio e C.Arrigo, Alessandria rboverio@ospedale.al.it
ASST Grande Ospedale Metropolitano Niguarda, Milano	Dott. Nicolò Capsoni Medico Medicina d'urgenza e pronto soccorso Ospedale Metropolitano Niguarda, Milano nicolo.capsoni@gmail.com
Ospedale Maggiore Carlo Alberto Pizzardi, Bologna	Dott. Alessio Bertini Direttore Medicina e Chirurgia d'Accettazione e Urgenza Ospedale Maggiore Carlo Alberto Pizzardi, Bologna alessio.bertini@yahoo.it
Ospedale Santa Maria Annunziata, Bagno a Ripoli (FI)	Dott.ssa Germana Ruggiano Direttore Medicina e Chirurgia d'Accettazione e Urgenza Ospedale Santa Maria Annunziata, Bagno a Ripoli (FI) germana.ruggiano@uslcentro.toscana.it
Ospedale Santa Maria delle Grazie, Pozzuoli (NA)	Dott. Giovanni Porta Medico Medicina d'urgenza e pronto soccorso Ospedale Santa Maria delle Grazie, Pozzuoli (NA) dottorgiovanniporta@gmail.com
Azienda sanitaria universitaria Friuli Centrale, Udine	Dott.ssa Gloria Zanello Direttore Medicina e Chirurgia d'Accettazione e Urgenza Azienda sanitaria universitaria Friuli Centrale, Udine mario.calci@asufc.sanita.fvg.it
Ospedale San Luigi Gonzaga, Orbassano	Dott.ssa Adriana Boccuzzi Medico Medicina d'urgenza e pronto soccorso Ospedale San Luigi Gonzaga, Orbassano valeria.caramello@yahoo.it
AOU Policlinico 'G.Rodolico – San Marco', Catania	Dott. Giuseppe Carpinteri Direttore Medicina e Chirurgia d'Accettazione e Urgenza PO Gaspare Rodolico, Catania gicarpinteri@gmail.com
Ospedale Sant'Eugenio, Roma	Dott.ssa Ersilia Castaldo Medico Medicina d'urgenza e pronto soccorso Ospedale Sant'Eugenio, Roma ersilia.castaldo@gmail.com
ASST Papa Giovanni XXIII, Bergamo	Dott. Roberto Cosentini Direttore Medicina e Chirurgia d'Accettazione e Urgenza ASST Papa Giovanni XXIII, Bergamo rcosentini@hpg23.it
Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano	Prof. Giorgio Costantino Direttore Medicina e Chirurgia d'Accettazione e Urgenza Ospedale Maggiore Policlinico, Milano giorgio.costantino@policlinico.mi.it
Ospedale San Martino, Genova	Dott.ssa Ombretta Cutuli Medico Medicina d'urgenza e pronto soccorso

	Ospedale San Martino, Genova ombretta.cutuli@hsanmartino.it
Ospedale degli Infermi, Rivoli	Dott.ssa Cristina Di Stefano Medico Medicina d'urgenza e pronto soccorso Ospedale degli Infermi, Rivoli cristinamaria.distefano@aslto3.piemonte.it
Ospedale Sant'Andrea, Vercelli	Dott. Sergio Tartaglia Direttore Medicina e Chirurgia d'Accettazione e Urgenza Ospedale Sant'Andrea, Vercelli srtartaglia@gmail.com
Ospedale Maggiore, Lodi	Dott. Stefano Paglia Direttore Medicina e Chirurgia d'Accettazione e Urgenza Ospedale Maggiore, Lodi Stefano.Paglia@asst-lodi.it

NLP-DeVal: Development and validation of a natural language processing tool to enable clinical research in emergency and acute care medicine: retrospective cohort study

ANNEX 2

Virtual case report form: dyspnea and transient loss of consciousness

Version 1.1

Ranica, February 7, 2024

eCREAM Coordinating Center
Istituto di Ricerche Farmacologiche Mario Negri IRCCS
Villa Camozzi. via GB Camozzi 3 - 24020 Ranica (Bergamo)

tel. +39 035 4535389 - fax +39 035 4535354 - e-mail: ecream@marionegri.it – Website ecreamproject.eu

1. **DYSPNEA**

1.1 Demographic

- Sex [male/female]
- Age [number]

1.2 History taking

- History of allergy [yes/no]
- History of recent trauma [yes/no]
- Pregnancy [yes/no]
- Presence of chronic pulmonary disease (either obstructive or restrictive) [yes/no]
- Chronic organ failure (respiratory, cardiac, renal, metabolic) [yes/no]
- Diffuse vascular disease [yes/no]
- Chronic rheumatologic disease [yes/no]
- Active neoplasia [yes/no]
- Chronic dialysis [yes/no]
- Immunosuppression [yes/no]
- Palliative care [yes/no]
- Neuropsychiatric disorders [yes/no]
- Poly-pharmacological therapy [yes/no]
- Problematic family context [yes/no]
- Need but absence of a caregiver [yes/no]
- Homelessness [yes/no]
- Living alone [yes/no]

1.3 Clinical examination

- Dementia [yes/no]
- General condition deterioration [yes/no]
- Level of autonomy (mobility) [4 levels: walking independently, walking with auxiliary aids, walking with physical assistance, bedridden]
- Level of consciousness [4 levels: AVPU]
- Agitation [yes/no]
- Presence of respiratory distress [yes/no]
- Foreign body in the airways [yes/no]
- Respiratory rate [number]
- Body temperature [number]
- Heart rate [number]
- Blood pressure [number]
- Improvement of dyspnea [yes/no]

1.4 Lab test results

- SpO_2 [number]
- pH [number]
- PaO_2 [number]
- $PaCO_2$ [number]
- HCO_3^- [number]
- Lactates [number]
- Hemoglobin [number]
- Platelets [number]
- Leukocytes [number]
- C-Reactive Protein [number]
- Blood sodium [number]
- Blood potassium [number]
- Blood glucose [number]
- Creatinine [number]
- Transaminases [number]
- INR [number]
- Troponin [number]
- BNP / NT-pro-BNP [number]
- D-dimer [number]
- SARS-CoV-2 swab test [pos/neg]

1.5 Imaging test results

- Thoracic ultrasound, any abnormalities [yes/no]
- ECG, any abnormalities [yes/no]
- Head/neck CT, any abnormalities [yes/no]
- Chest CT, any abnormalities [yes/no]
- Chest Rx, any abnormalities [yes/no]
- Gastroscopy [yes/no]

1.6 Treatment

- Performance of thoracentesis [yes/no]
- Administration of diuretics [yes/no]
- Administration of steroids [yes/no]
- Administration of bronchodilators [yes/no]
- Administration of oxygen/ventilation [yes/no]
- Blood transfusions [yes/no]

1.7 Final diagnosis

- Situational syncope [yes/no]

- Pulmonary embolism [yes/no]
- Heart failure [yes/no]
- Pneumonia [yes/no]
- COPD exacerbation [yes/no]
- Acute pulmonary edema [yes/no]
- Asthma exacerbation [yes/no]
- Severe Anemia [yes/no]
- Arrhythmia [yes/no]
- Respiratory failure [yes/no]
- Intoxication [yes/no]
- COVID 19 [yes/no]
- Influenza and various infections [yes/no]
- Pneumothorax [yes/no]
- Acute coronary syndrome [yes/no]

1.8 Follow-up

- 30-day mortality [dead/alive] (if possible)

2. **TRANSIENT LOSS OF CONSCIOUSNESS**

2.1 Demographic

- Sex [male/female]
- Age [number]

2.2 History taking

- History of drug abuse [yes/no]
- History of alcohol abuse [yes/no]
- Presence of pacemaker [yes/no]
- Presence of defibrillator [yes/no]
- Cardio-pulmonary resuscitation [yes/no]
- Antihypertensive therapy [yes/no]
- Anticoagulants or antiplatelet drug therapy [yes/no]
- Diagnosis of cardiovascular diseases [yes/no]
- Diagnosis of neurodegenerative diseases [yes/no]
- Diagnosis of peripheral neuropathy [yes/no]
- Pregnancy [yes/no]
- Diagnosis of Chronic respiratory failure [yes/no]
- Diagnosis of Chronic cardiac failure [yes/no]
- Diagnosis of Chronic renal failure [yes/no]
- Diagnosis of Chronic metabolic failure [yes/no]

- Diagnosis of Diffuse vascular disease [yes/no]
- Diagnosis of Chronic rheumatologic disease [yes/no]
- Active neoplasia [yes/no]
- Chronic dialysis [yes/no]
- Immunosuppression [yes/no]
- Palliative care [yes/no]
- Neuropsychiatric disorders [yes/no]
- Poly-pharmacological therapy [yes/no]
- Speed with which the patient recovered consciousness [2 levels: fast (seconds)/slow (minutes-hours)]
- Presence of prodromal symptoms [yes/no]
- Compliance with antiepileptic therapy [yes/no]
- Duration of the patient's unconsciousness [2 levels: short (seconds)/long (minutes-hours)]
- TLoC during effort [yes/no]
- TLoC while supine [yes/no]
- First episode vs. known history of epilepsy [yes/no]
- Antiepileptic therapy already in place [yes/no]
- Drowsiness, confusion, disorientation as postcritical state [yes/no]
- Pale skin during the episode [yes/no]
- Eye deviation during the episode [yes/no]
- Stiffness during the episode [yes/no]
- Drooling during the episode [yes/no]
- Tonic-clonic seizures [yes/no]
- Situation description, like coughing, prolonged periods of straining, sudden abdominal pain, phlebotomy [yes/no]
- Problematic family context [yes/no]
- Need but absence of a caregiver [yes/no]
- Homelessness [yes/no]
- Living alone [yes/no]

2.3 Clinical examination

- Dementia [yes/no]
- General condition deterioration [yes/no]
- Level of autonomy (mobility) [4 levels: walking independently, walking with auxiliary aids, walking with physical assistance, bedridden]
- Level of consciousness [4 levels: AVPU]
- Agitation [yes/no]
- Presence of dyspnea [yes/no]
- Body temperature [number]
- Heart rate [number]
- Blood pressure [number]

- Chest pain [yes/no]
- Head or other districts trauma [yes/no]
- Ab ingestis pneumonia [yes/no]
- Tongue bite [yes/no]
- Further seizures in the ED [yes/no]
- Blood in the stool [pos/neg]
- Carotid sinus massage [yes/no]
- Supine-to-standing systolic blood pressure test [yes/no]
- Improvement of patient's conditions [yes/no]
- Neurologist consultation [yes/no]

2.4 Diagnostic test results

- ECG, any abnormality [yes/no]
- ECG monitoring, any abnormality [yes/no]
- EEG, any abnormality [yes/no]

2.5 Lab test results

- SpO_2 [number]
- Lactates [number]
- Hemoglobin [number]
- Platelets [number]
- Leukocytes [number]
- C-Reactive Protein [number]
- Blood glucose [number]
- Blood sodium [number]
- Blood potassium [number]
- Blood calcium [number]
- Creatinine [number]
- Transaminases [number]
- INR [number]
- Troponin [number]
- BNP / NT-pro-BNP [number]
- D-Dimer [number]
- Serum creatine kinase [number]
- Blood alcohol [number]
- Blood drug dosage [number]
- Urine drug test [number]

2.6 Imaging test results

- Brain CT scan, any abnormality [yes/no]

- Brain MRI, any abnormality [yes/no]
- Cardiac ultrasound, any abnormality [yes/no]
- Chest CT scan, any abnormality [yes/no]
- Pulmonary scintigraphy, any abnormality [yes/no]
- Gastroscopy [yes/no]
- Abdomen CT scan, any abnormality [yes/no]
- Doppler ultrasound of deep veins, any abnormality [yes/no]
- Compression ultrasound (CUS), any abnormality [yes/no]

2.7 Treatment

- Administration of fluids [yes/no]

2.8 Final diagnosis

- Situational syncope [yes/no]
- Epilepsy / epileptic seizure [yes/no]
- Pulmonary embolism [yes/no]
- Arrhythmia [yes/no]
- Cardiac tamponade [yes/no]
- Aortic dissection [yes/no]
- Acute coronary syndrome [yes/no]
- Hemorrhage [yes/no]
- Severe Anemia [yes/no]
- Concussive head trauma [yes/no]

2.9 Follow-up

- 30-day mortality [dead/alive] (if possible)

Development of a natural language processing tool to enable clinical research in emergency and acute care medicine: retrospective cohort study

ANNEX 3

AnonymAI accuracy report

eCREAM Coordinating Center
Istituto di Ricerche Farmacologiche Mario Negri IRCCS
Villa Camozzi. via GB Camozzi 3 - 24020 Ranica (Bergamo)

tel. +39 035 4535389 - fax +39 035 4535354 - e-mail: ecream@marionegri.it – Website ecreamproject.eu

maize.
A JAKALA COMPANY

Spett.le
Fondazione Bruno Kessler
Via Santa Croce, 77
38122 Trento
Italia

Oggetto: Report accuratezza AnonymAI per progetto Horizon EU eCREAM

Introduzione

Il nostro strumento di anonimizzazione nasce come un progetto di ricerca finanziato dalla Unione Europea nella terza call EU “NGI Trust” (Next Generation Internet).

Il nostro anonimizzatore è stato progettato per proteggere la privacy degli utenti attraverso l'anonimizzazione dei dati personali. Per lo sviluppo dello strumento abbiamo raccolto sei corpora per un totale di oltre 57.000 documenti contenenti poco più di 94.000 annotazioni. I corpora riguardano diversi ambiti, tra cui trascrizioni di sessioni paziente-terapista, recensioni del servizio clienti bancario ed e-mail aziendali. Il dataset comprende documenti in due lingue, italiano e inglese, rappresentati rispettivamente da circa 61.500 e 32.500 annotazioni.

AnonymAI utilizza una varietà di algoritmi di intelligenza artificiale, tra cui Machine Learning, Deep Learning e sistemi a regole, al fine di identificare le parti di un documento di testo che contengono informazioni personali. Una volta individuate, queste sezioni vengono rimpiazzate da etichette anonime nel documento, che viene restituito al mittente privo di tali contenuti sensibili.

Questo report documenta l’accuratezza attesa di tale strumento su documenti di natura generica attraverso le metriche presentate di seguito.

Metriche Entità Singole

La valutazione dello strumento è stata effettuata su un set di test separato che non era stato utilizzato nella fase di training e non era stato visto dall'algoritmo. La seguente tabella riporta le metriche relative al riconoscimento delle entità facenti parte del progetto eCREAM.

Tabella 1: le prestazioni dello strumento divise per entità da anonimizzare.

Entità	Pos	Precision	Recall	F1 score
Nome	B	0.964	0.982	0.973
Cognome	B	0.947	0.975	0.961
Numero di cellulare	B	1	0.98	0.99
Città, provincia, comune	B	0.94	0.96	0.95
	I	0.915	0.932	0.924
Via, corso, indirizzo	B	0.947	0.989	0.968
	I	0.915	0.985	0.949
Relazione di Parentela	B	0.926	0.959	0.942

La colonna "Pos" mostra la posizione del token (parola o parte di essa) all'interno della sequenza di parole analizzate che fanno parte dell'entità. 'B' significa che il token è l'inizio di una sequenza di interesse, mentre 'I' significa che si trova all'interno di una sequenza. Ad esempio, per l'entità "*via San Quintino*":

via → *indirizzo-B*
San → *indirizzo-I*
Quintino → *indirizzo-I*

"Precision" misura la percentuale di identificazioni di entità che sono effettivamente corrette. Corrisponde al valore predittivo positivo dello strumento. "Recall" misura la percentuale di entità presenti nel testo che sono identificate correttamente. Corrisponde alla sensibilità dello strumento.
"F1 score" è una misura che combina precisione e recall in un unico numero, fornendo una valutazione sintetica dell'accuratezza dello strumento.

Il valore F1-score, che è una media armonica tra Precision e Recall, è notevolmente alto per tutte le categorie di dati coinvolte nel progetto, con il valore più basso registrato per città/provincia/comune (95%) e i più alti per i nomi (97.3%) e i numeri di cellulare (99%).

Questi risultati dimostrano che il nostro strumento di anonimizzazione è in grado di operare con precisione e affidabilità, garantendo un alto livello di protezione per un'ampia varietà di dati sensibili.

Si fa notare che, per quanto riguarda le esigenze del progetto eCREAM, la metrica di interesse è “Recall”, ovvero la probabilità che le occorrenze di una determinata entità siano identificate correttamente (e di conseguenza anonimizzate), sul totale di quelle presenti nei testi da anonimizzare.

Metriche Entità Combinate

Le maggiori criticità in termini di protezione dei dati personali si verificano qualora siano disponibili più informazioni su uno stesso soggetto, aumentando in maniera significativa la possibilità di reidentificazione da un lato (es. nome, cognome, indirizzo) o di conoscere di un soggetto alcuni dati personali (es. sapere che un soggetto identificato ha avuto un parente trattato in Pronto Soccorso). Si ricorda infatti che il progetto eCREAM intende analizzare testi liberi presenti nelle cartelle cliniche di Pronto Soccorso (PS) e che in tale testi è possibile che compaiano riferimenti a persone terze (tipicamente un parente del paziente) con le modalità per contattarlo (ad es. il numero di telefono).

Unendo le metriche di accuratezza sulle entità singole, possiamo calcolare le probabilità che combinazioni di entità non vengano anonimizzate correttamente dallo strumento AnonymAI.
La seguente tabella riporta le probabilità combinate per le entità di interesse nel progetto eCREAM:

Tabella 2: Probabilità che varie combinazioni di entità non vengano anonimizzate.

Combinazione identificativa	Probabilità di NON anonimizzare un soggetto in una nota	Numero di note da analizzare per avere un soggetto NON anonimizzato
Nome completo (nome + Cognome)	0.00045	2,222
Nome completo + numero di telefono	0.000009	111,111
Nome completo + indirizzo completo	1.98 E-7	5,050,505
Nome completo + numero di telefono + indirizzo completo	3.96 E-9	252,525,253
Nome completo + grado di parentela	1.85 E-5	54,054
Nome completo + numero di telefono + grado di parentela	3.69 E-7	2,710,027
Nome completo + indirizzo completo + grado di parentela	8.118 E-9	123,183,050
Nome completo + numero di telefono + indirizzo completo + grado di parentela	1.6 E-10	6,250,000,000

maize.

A JAKALA COMPANY

La tabella elenca le varie combinazioni delle entità da anonimizzare nel progetto eCREAM. Per ogni combinazione, riporta la probabilità che il nostro strumento non anonimizzi un soggetto che è invece menzionato nella nota (seconda colonna), nonché il numero di note contenenti il riferimento a un soggetto che devono essere analizzate per averne una in cui il soggetto non viene anonimizzato dal nostro strumento (terza colonna).

Ad esempio, la probabilità di non riuscire a rilevare contemporaneamente sia il nome sia il cognome di una persona citata in un testo libero risulta pari a 0.00045. Ciò significa che il nostro strumento deve analizzare mediamente 2222 testi in cui è riportato nome e cognome di un individuo prima di non riuscire a rilevare una tale combinazione e quindi di non riuscire ad anonimizzarla.

La tabella mostra anche quale sia la probabilità non solo di non riuscire ad anonimizzare un individuo, ma anche di non riuscire a togliere l'informazione, una volta identificato, relativa al suo (eventuale) grado di parentela con il paziente a cui si riferisce il testo. Si tratta di probabilità sempre estremamente basse.

In conclusione, i dati mostrano che il nostro strumento di anonimizzazione possiede un'alta accuratezza. Per le combinazioni di informazioni a più alto rischio di rivelare l'identità di una persona e la sua relazione con un paziente di Pronto Soccorso, la probabilità di non essere anonimizzate è matematicamente trascurabile, confermando l'efficacia del nostro strumento nel proteggere la privacy e l'anonimato degli utenti.

MAIZE S.r.l

NLP-DeVal: Development and validation of a natural language processing tool to enable clinical research in emergency and acute care medicine: retrospective cohort study

ANNEX 4

Technical aspects of data processing and protection

Version 1.1

Ranica, February 7, 2024

eCREAM Coordinating Center
Istituto di Ricerche Farmacologiche Mario Negri IRCCS
Villa Camozzi. via GB Camozzi 3 - 24020 Ranica (Bergamo)

tel. +39 035 4535389 - fax +39 035 4535354 - e-mail: ecream@marionegri.it – Website ecreamproject.eu

1. DATA TRANSMISSION

1.1. Exposure of APIs on the local eCREAM platform

Integration APIs between hospital systems and the *Local eCREAM NLP platform* will be exposed on the intranet, not on the internet. Thus, transmission of data that has not yet been anonymized will take place over the hospital's internal network.

The integration APIs between local eCREAM and the *Central eCREAM NLP server* will be exposed on the Internet, the anonymized data will be encrypted and sent to the NLP server.

2. AUTHENTICATION

In order to invoke one or more APIs, the ED system must be an application surveyed by eCREAM, which will receive from eCREAM the authentication credentials needed by the application in the Access Token request phase; this Token will have to be attached to each subsequent API invocation.

The application's credentials must not be disclosed or reused to build other applications. Each new application must be individually registered so that it will be issued the corresponding credentials.

It is the responsibility of the user to properly protect the credentials used by the various application instances.

eCREAM will verify that the credentials used for each integration API access with the local eCREAM system are associated with the previously surveyed hospital network.

Regarding access to the integration API with the NLP server, the eCREAM system will verify, with each connection, that the credentials are associated with a certificate issued by eCREAM and installed on the calling machines. The Access Token is obtained by calling a special API in the eCREAM system. An Access Token has a limited life of 30 minutes so, when it expires, the Token must be renewed through the appropriate API of the eCREAM system.

Each requester will need to authenticate and request permission to send previously anonymized data or to request anonymized data currently present on the *Central eCREAM NLP server*.

2.1. Authorization

Access to the API will be subject to an authorization system that leverages the authentication features described above.

API access will be associated with compartmentalization criteria called scope.

When a hospital application is censored in eCREAM, the roles required to use the corresponding scopes will be associated.

When issuing the Access Token, the Authentication API will require the caller to specify the list of scopes associated with the API/resources it will use through that Token, or else authorization in accessing the API will be denied at the point of fruition.

2.2. Logging and Tracking API Accesses

All interactions between the application and external systems will generate a record in the application log.

The log will not contain sensitive data, but only data related to system operations or incidents captured by the system.

It will allow identification of the calling user or system, the operation with the metadata necessary to verify

whether the operation in question affected a specific dataset of sensitive data, and the timestamp of the operation.

This will make it possible to identify any abusive access or attempted access to data.

3. DATA PROCESSING IN THE LOCAL ECREAM NLP-DEVAL PLATFORM

The components that make up the *Local eCREAM NLP platform* will be entirely installed within the server farms operated by the hospital entities.

The local eCREAM component will receive sensitive data associated with subject identification data contained in free text fields and clinical documents.

The data processing procedure involves transmitting the data to the local eCREAM component, which will provide further processing and subsequent transmission, in anonymized form, to the *Central eCREAM NLP server*. The data will be transferred to the server through secure protocols.

3.1. Anonymization

Data from the hospital EHRs will be anonymized within the hospital premises, with the use of the AnonymAI software. This software program will remove data entities referring to specific types of personal data from free text, specifically: first name, last name, phone number, address, location, and parental relationship. Once anonymized, the data will be transferred to the central NLP server.

3.2. Storage and retention of anonymized data within the Local eCREAM NLP platform

Databases and file systems that store the anonymized data within the *Local eCREAM NLP platform* will be encrypted. Data will be retained in these archives only as long as necessary for their further processing and subsequent transmission to the *Central eCREAM NLP server*. At the end of the transmission process, the data will be physically deleted.

3.3. Physical access control

With regard to the local eCREAM component, physical access control will be the responsibility of the hospital entity, the data controller.

3.4. Logical access control

Data access will be allowed only to system administrators. Control will be performed by configuring the enabled connection IPs and verifying the validity of the required certification associated with the individual enabled IP.

Access will be conveyed via VPN and allowed only through personal credentials, provided only to administrators appointed to process the data.

3.5. Logging, tracking of access and activities performed

All interactions between the application and external systems, including system administrators, will generate a record in the application log.

The log will not contain sensitive data, but only data related to operations performed through the system or incidents captured by the system.

It will allow identification of the calling user or system, the operation with metadata necessary to verify whether the operation in question affected a specific dataset of sensitive data, and the timestamp of the operation.

This will help identify any abusive access or attempts to access data.

3.6. Antivirus

The local eCREAM system will allow files to be transmitted to them, so they will use a server-side antivirus to scan incoming files.

4. DATA PROCESSING IN THE CENTRAL ECREAM NLP SERVER

4.1. Data storage

Databases and file systems will be encrypted.

4.2. Data retention

The anonymized free texts that will have been transmitted to the NLP server, located in Italy at the IRFMN in Milan, will not fall within the scope of the GDPR. These anonymized free texts will be transmitted to the Fondazione Bruno Kessler, to Orobix Life, and to the main European platforms for language resource sharing, including the European Language Grid repository (https://live.europeanlanguage-grid.eu), the CLARIN Virtual Language Observatory (https://vlo.clarin.eu/?2), and the European Language Equality (ELE) initiative (https://european-language-equality.eu) and will be usable by users worldwide.

4.3. Backup

The processed data present in the central NLP server will be backed up daily via a fully automatic process, with a 30-day data retention period.

Backup databases and file systems will be encrypted.

4.4. Physical access control

The data present on the NLP server will be processed locally. Access is controlled, registered, and limited to those with access permission. The server farm room is monitored for safety and prevention (UPS, temperature, theft, fire).

Physical access control for the *Central eCREAM NLP server* will be the responsibility of IRFMN, the data controller.

4.5. Logical access control

Data access to the central NLP server will be allowed only to system administrators. Control will be performed by configuring the enabled connection IPs and verifying the validity of the required certification associated with the individual enabled IP.

Access will be allowed on a nominal basis to system administrators, authorized only through an access machine having a registered certificate and authorized IP.

4.6. Logging, tracking of access and activities performed

Concerning the data on the NLP server, all interactions between the application and external systems, including system users, will generate a record in the application log.

The log will not contain sensitive data, but only data related to operations performed through the system or incidents captured by the system.

It will allow identification of the calling user or system, the operation with metadata necessary to verify whether

the operation in question affected a specific dataset of sensitive data, and the timestamp of the operation.

This will help identify any abusive accesses or attempts to access data.

4.7. Antivirus

The central NLP server will use a server-side antivirus that keeps the delivery environment scanned.