eCream meeting – Košice
Oct 2026

De-identification of clinical notes

Now you see me, now you don’t!
Who wins the anonymization game in healthcare?

Pietro Ferrazzi – NLP Unit @ FBK

FONDAZIONE BRUNO KESSLER

How to use

1 **NLP-DeVal**
What we said we should do

2 **The process in place**
What are the steps the notes have to go through

3 **The Problems**
FP

4 **The Shared Task**
How to open to the research community

How to use

1 **NLP-DeVal**
What we said we should do

2 **The process in place**
What are the steps the notes have to go through

3 **The Problems**
FP

4 **The Shared Task**
How to open to the research community

Stuff we are happy about

Stuff we are *less* happy about

The eCREAM Standard

eCREAM
enabling Clinical Research in Emergency and Acute care Medicine

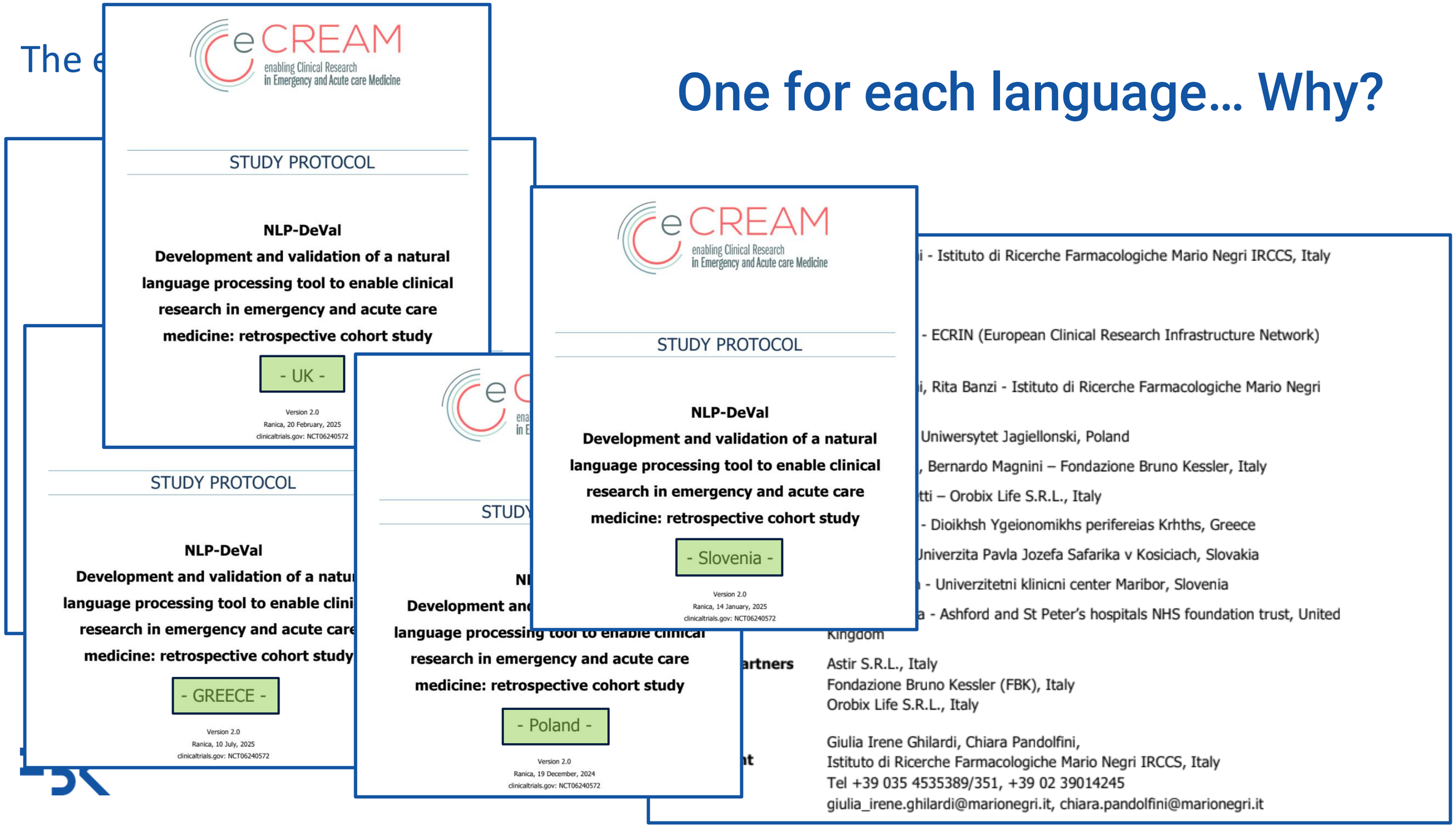
STUDY PROTOCOL

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

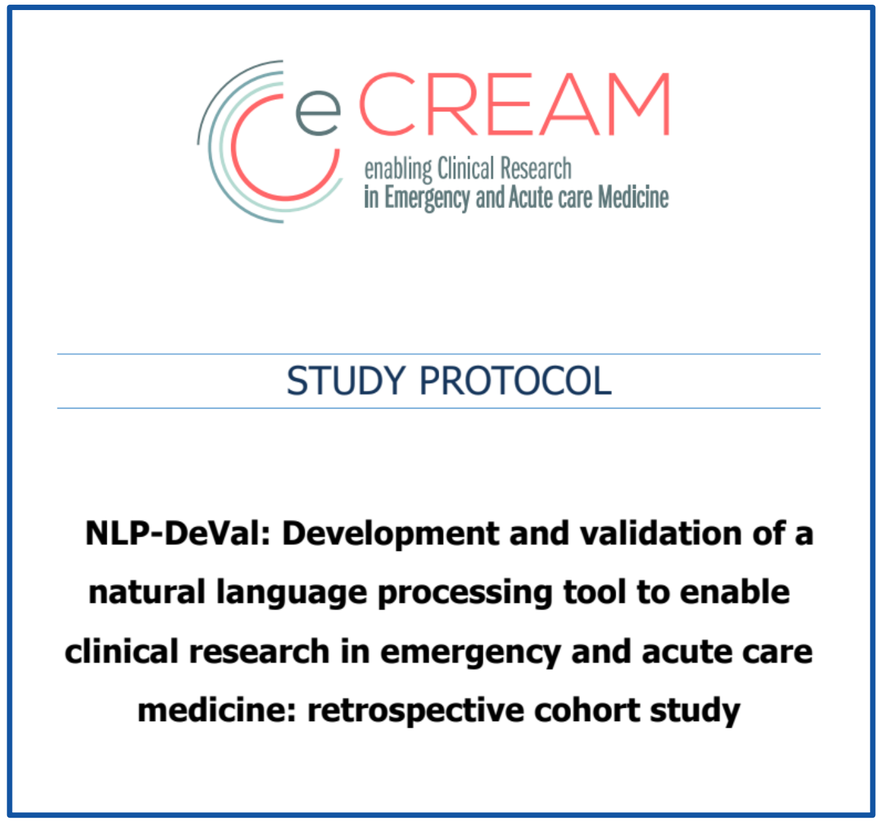
The Project has defined a Study Protocol submitted to the Ethical Committees of ED

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

One for each language… Why?



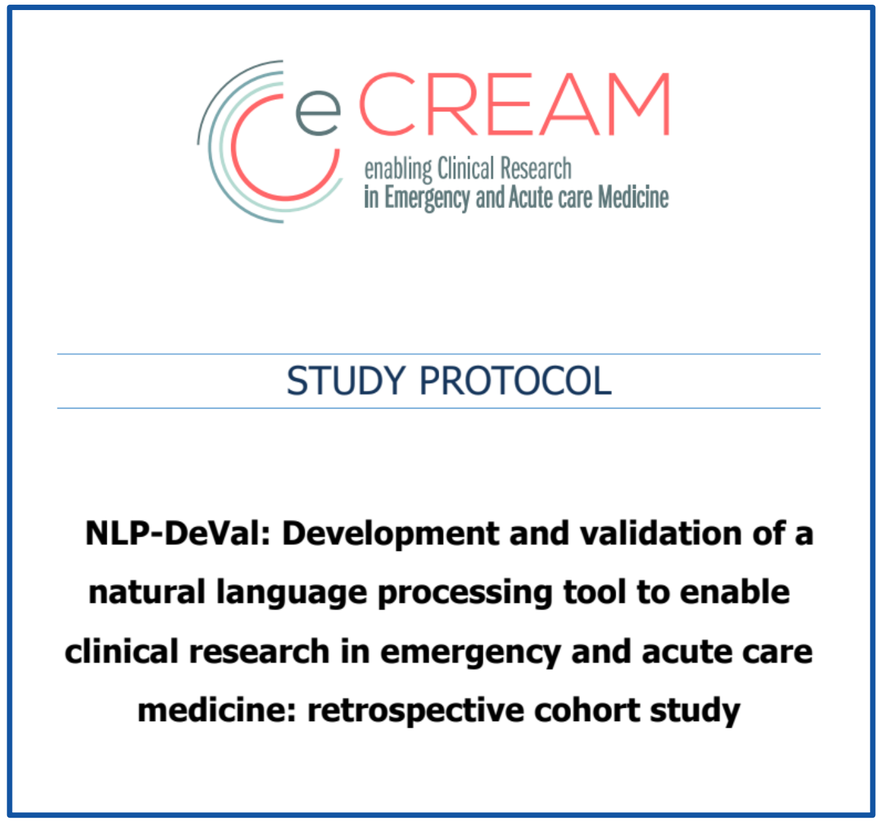
The eCREAM Standard



There are two main Sections that define how anonymization can carried out inside eCREAM:

- There is an explicit point on *"Data collection and anonymization"* (5.3)
- A fundamental part is contained at *"The legal basis for data processing"* (8.1)

The eCREAM Standard

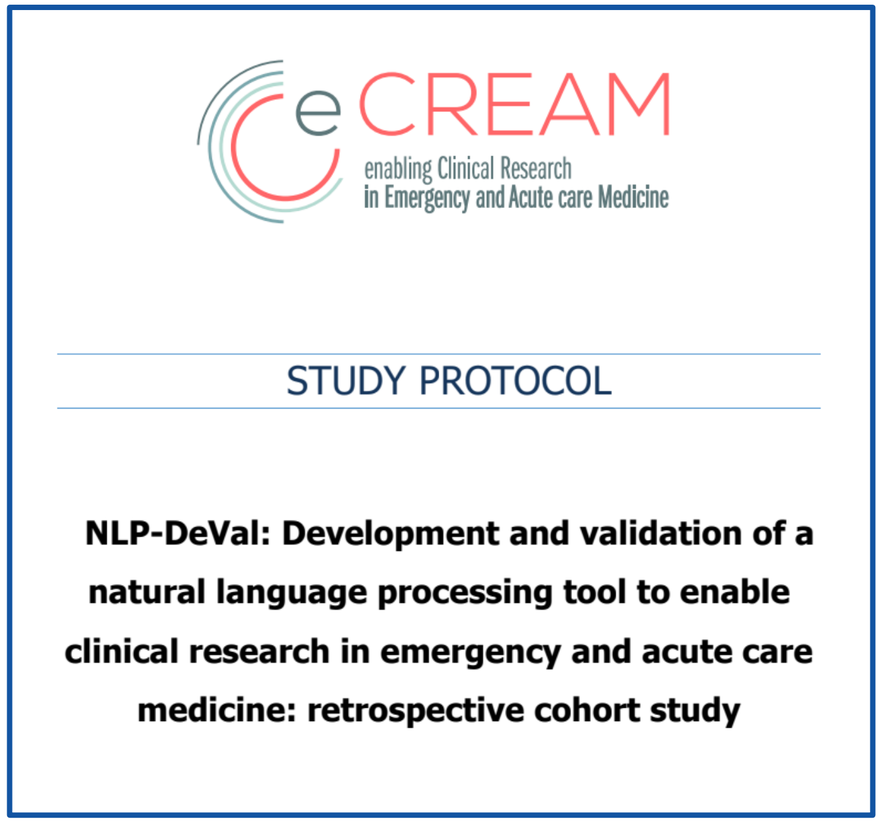


There are two main Sections that define how anonymization can carried out inside eCREAM:

This is language-dependent!

- There is an explicit point on *"Data collection and anonymization"* (5.3)
- A fundamental part is contained at *"The legal basis for data processing"* (8.1)

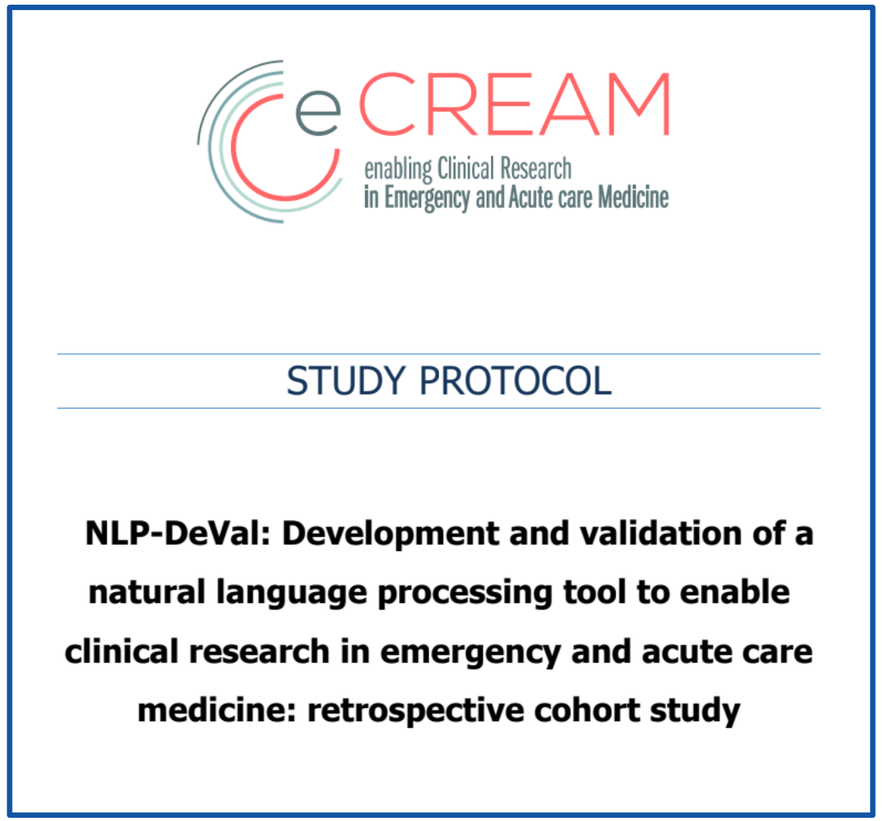
The eCREAM Standard



There are n elements that can identify a patients:

1. Explicit information about them, or the context of their stay at the hospital

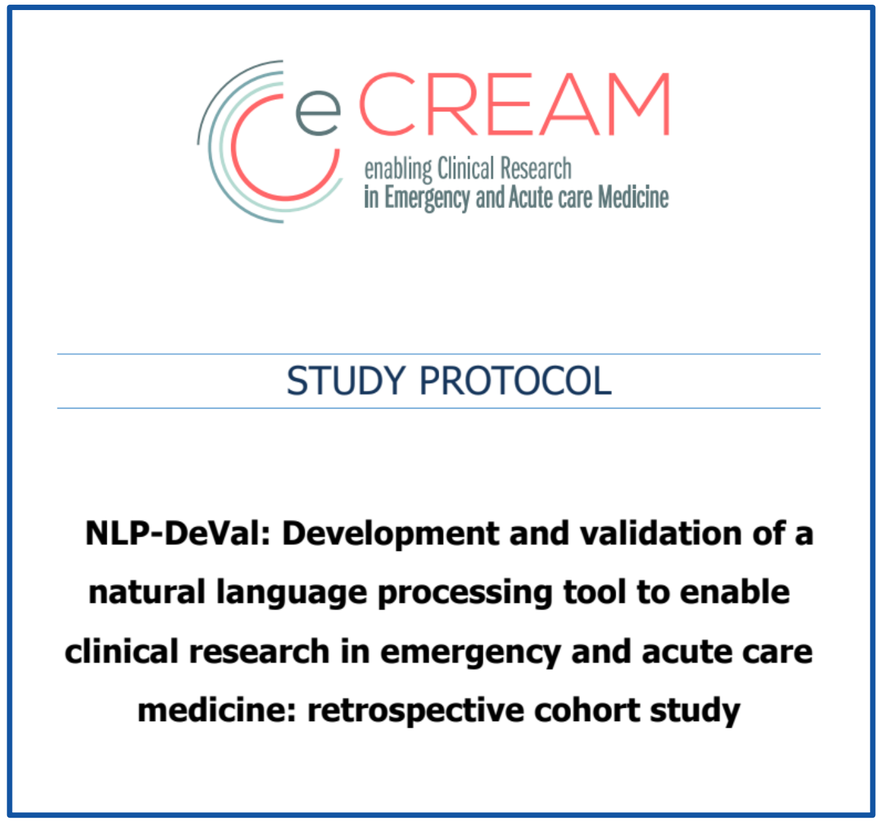
The eCREAM Standard



There are n elements that can identify a patients:

1. Explicit information about them, or the context of their stay at the hospital
2. The combination of several steps involved in a long medical history

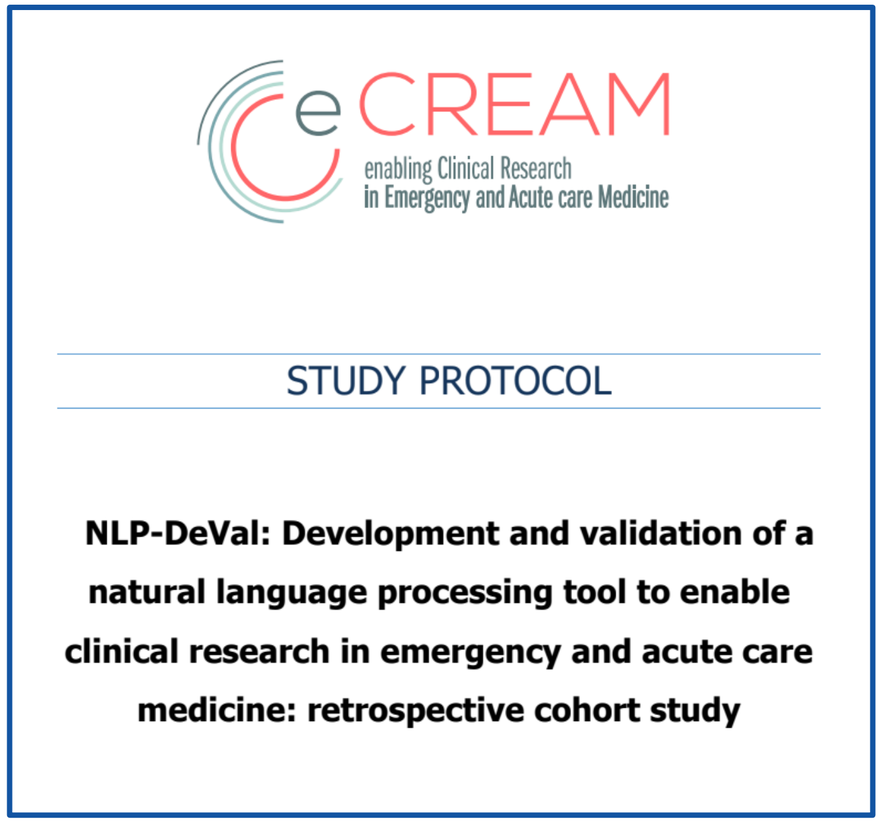
The eCREAM Standard



There are n elements that can identify a patients:

1. Explicit information about them, or the context of their stay at the hospital
2. The combination of several steps involved in a long medical history
3. If the reader knows already a lot about a patient, they could find something new, especially if they have some special conditions

The eCREAM Standard



There are n elements that can identify a patients:

1. Explicit information about them, or the context of their stay at the hospital
2. The combination of several steps involved in a long medical history
3. If the reader knows already a lot about a patient, they could find something new, especially if they have some special conditions
4. The information about other people, involved in the caring process

The eCREAM Standard

To be taken care of

Fully convered by the design choices

To be taken care of

1. Explicit information about them, or the context of their stay at the hospital
2. The combination of several steps involved in a long medical history

   If the reader knows already a lot about a patient, they could find something new, especially if they have some special conditions
3.
4. The information about other people, involved in the caring process

The eCREAM Standard

Now we will briefly se how this is embedded in the process,



Fully convered by the design choices

2 ● The combination of several steps involved in a long medical history

3 ● If the reader knows already a lot about a patient, they could find something new, especially if they have some special conditions

The eCREAM Standard

Now we will briefly se how this is embedded in the process,

Then, our approach to this

Fully convered by the design choices

2. The combination of several steps involved in a long medical history

3. If the reader knows already a lot about a patient, they could find something new, especially if they have some special conditions

To be taken care of

1. Explicit information about them, or the context of their stay at the hospital

To be taken care of

4. The information about other people, involved in the caring process

Fully convered by the design choices

Dimensions to take care of I/IV

The combination of several steps involved in a long medical history

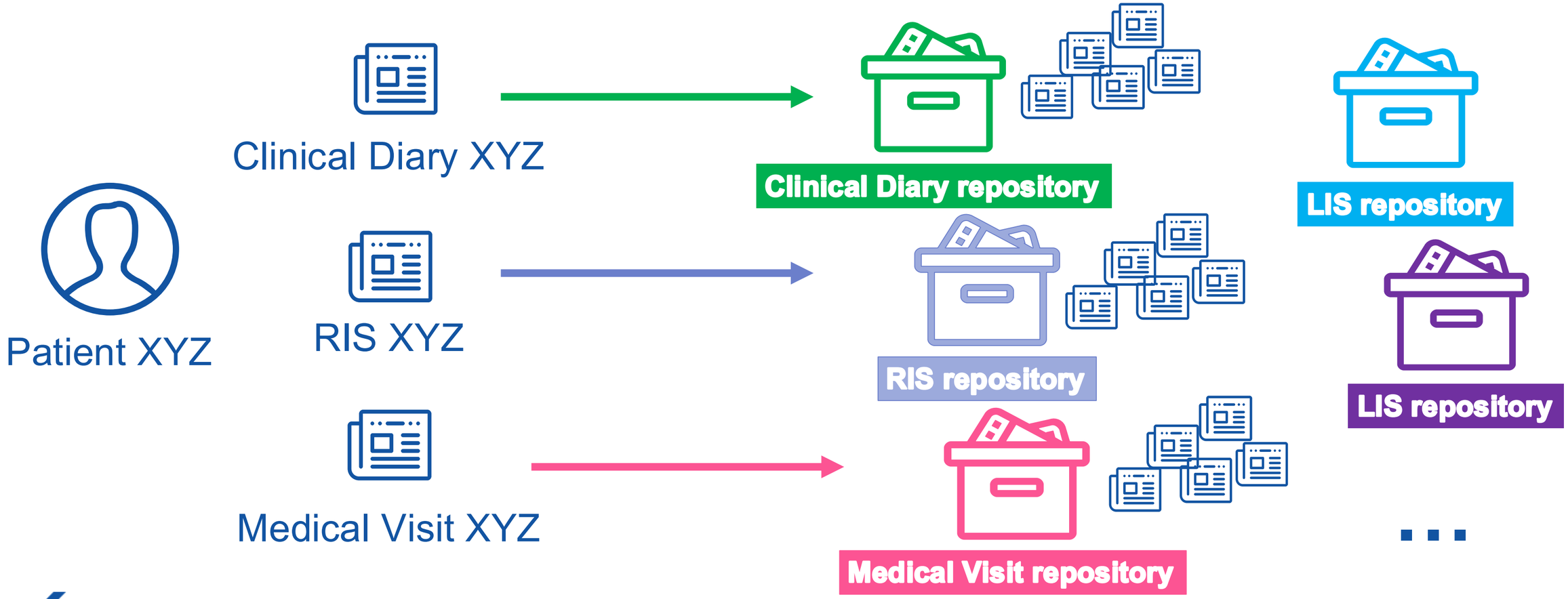
> “the notes referring to different aspects of the same patient […] will be separated from each other so that it will be impossible to reconstruct the entire patient profile”



Dimensions to take care of I/IV

The combination of several steps involved in a long medical history

Fully convered by the design choices



Dimensions to take care of II/IV

Fully convered by the design choices

If the reader knows already a lot about a patient, they could find something new, especially if they have some special conditions

> *“to identify a specific patient in a database, one should already know all the characteristics that make that patient unique in that database. […] 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”*

Dimensions to take care of II/IV

Fully convered by the design choices

If the reader knows already a lot about a patient, they could find something new, especially if they have some special conditions

> *to identify a specific patient in a database, one should already know all the characteristics that make that patient unique in that database. […] 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*

> *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.*

Dimensions to take care of II/IV

Fully convered by the design choices

If the reader knows already a lot about a patient, they could find something new, especially if they have some special conditions

> *to identify a specific patient in a database, one should already know all the characteristics that make that patient unique in that database. […] 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*

> *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.*

And we have got hundred of thousands!

To be taken care of

Dimensions to take care of III/IV

Explicit information about the patient, or their stay at the hospital

> “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.).”

This information lives in specific places: the headers!

To be taken care of

Dimensions to take care of III/IV

Explicit information about the patient, or their stay at the hospital

HEADER

Veneto, Padova - Azienda Sanitaria Locale Città di Padova-PD - Direttore Dott. Mario Rossi
Presidi Ospedalieri ABCD – Corso Milano
Tel. 800 1234567 - +39 3216548709 (Segreteria)- Mail: example@for_the_presentation.com
Risultati validati dall'operatore tecnico sanitario nel pieno rispetto delle procedure e dei protocolli operativi predisposti dal Direttore della S.C. Lab. Analisi

Esame	Esito	U.M.	Val. Riferimento	Metodica
Sg-Emocromo				
WBC (Leucociti)	13.95 >	x 10^3 / ul	4.00 - 10.00	
RBC (Eritrociti)	5.24	x 10^6 / ul	4.50 - 6.00	
Monociti val.assoluto	0.96	x 10^3 / ul		

[...]

S-Lipasi	34	U/l	13 - 60	
S-Proteina C reattiva	2.3 >	mg/dl	0.1 - 0.5	
S-Troponina T	0.04	ng/mL	< 0.03	

Troponina T eseguita su campione emolizzato, possibili interferenze analitiche. Ripetere prelievo.
Legenda Materiali: S = Siero P = Plasma Sg = Sangue U = Urine dU = Diuresi Oraria La= Liquidi Altri LCR = Liquor F = Feci

To be taken care of

Dimensions to take care of IV/IV

The information about other people, involved in the caring process

PROBLEM:

> “The procedures described above, [...], 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.”

To be taken care of

Dimensions to take care of IV/IV

The information about other people, involved in the caring process

PROBLEM:

> “The procedures described above, […], 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.”

PROPOSED SOLUTION:

> “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**.”

To be taken care of

Dimensions to take care of
Summary of open problems

Are we good at finding the headers?

Are we good at removing names, locations, ecc. ?

To be taken care of

Dimensions to take care of

Summary of open problems

Are we good at finding the headers?



Are we good at removing names, locations, ecc. ?



Quick Wrap-Up

- We have a path set for de-identification
  - Some parts come «for free»
  - Some parts don’t

Quick Wrap-Up

- We have a path set for de-identification
  - Some parts come «for free»
  - Some parts don’t

Let’s see what we can (are) doing

Anonimization tool

We commit to the AnonimAI, a tool from maize. A JAKALA COMPANY

Anonymization tool started out as a research project funded by the EU “NGI Trust”

Developed over 6 corpora (57,000 docs), on 94,000 annotations, in Italian (2/3) and English (1/3).
Focused on patient-therapist session transcripts, bank customer service reviews, company emails.

AnonymAI uses
- *Machine Learning,*
- *Deep Learning, and*
- *rule systems, in order to identify the parts of a text document that contain personal information.*

Once identified, these sections are replaced by anonymous labels in the document, which is returned to the sender without such sensitive content.

Anonimization tool

Only!?

Developed over 6 corpora (57,000 docs), on 94,000 annotations, in Italian (2/3) and English (1/3).

Anor



ns, in Italian (2/3) and English (1/3).

Only!?

Anonimization tool

“ only 10% of them will contain a reference to a third party ”

We create an eCream **test dataset** to verify if the anonimization tool is good enough on our settings

Entity	F1 score
First name	0.973
Surname	0.961
Mobile Phone number	0.99
City, provence, municipality	0.95
	0.924
Street, avenue, address	0.968
	0.949
Family relationship	0.942

Anonimization tool

We c
datas
anon
enou

Is it really a problem to find the word «mum» in it? Or a city? Or even a street?

Entity	F1 score
Family relationship	0.942

Anonimization tool

We c
datas
anon
enou

Entity	F1 score
Family relationship	0.942

> ***“we need to concentrate on the probability of not anonymizing the combination”***

Anonimization tool

Remember: this already excludes 90% of our notes!

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

Anonimization tool

All good until... we find more mistakes than expected

Descrizione Analisi	Risultato	U. di Misura	Int. di Riferimento	Metodo
V-Biochimica				
Procalcitonina	0.4	microg/L		
Chemiluminescenza				

valori < 0.5 microg/L indicano un BASSO rischio di sepsi grave e/o shock settico
valori >2.0 microg/L indicano un ALTO rischio di sepsi grave e/o shock settico

Per gi esami eseguiti presso il Laboratorio Di Novara,
responsabile della Validazione: Prof. NOME_PERSONA.

Our plan for an extra layer of de-identification

Exploratory analysis of the anonymized notes

Our plan for an extra layer of de-identification

Exploratory analysis of the anonymized notes

1. selected 500 notes from 2 hospitals (SGB, SA)
2. manually read them

hospital	length_class	id	category	word_length	source_file	text	TP	FP	FN	PARENTE_TP	PARENTE_FP	PARENTE_FN	NOME_PERSONA_TP	NOME_PERSONA_FP	NOME_PERSONA_FN	LUOGO_TP	LUOGO_FP	LUOGO_FN
					_DIARY/CLINICAL_DIARY_0.zip:1828396.txt	PARENTE												
vercelli	3	610165	NURSING_CARE_NOTES	16	Vercelli/NURSING_CARE_NOTES/NURSING_CARE_NOTES_0.zip:610165.txt	SI POSIZIONA IN LUOGO IN DEGENZA TEMPORANEA COVID FREE.VIGILE,ORIENTATO E COLLABORANTE.SI ESEGUE TERAPIA COME DA PRESCRIZIONE											1	
vercelli	3	953398	CLINICAL_DIARY	15	Vercelli/CLINICAL_DIARY/CLINICAL_DIARY_0.zip:953398.txt	allertata rianimazione di LUOGO la paziente viene trasferita presso l'ospedale di LUOGO con assistenza rianimatoria										2		
vercelli	3	1004077	DISCHARGE	38	Vercelli/DISCHARGE/DISCHARGE_0.zip:1004077.txt	TRAUMA CONTUSIVO-DISTORSIVO BASE DEL V DITO MANO DS DA CAUSA ACCIDENTALE A SCUOLA DURANTE ORA DI EDUCAZIONE FISICA SEGUIRE LE INDICAZIONI FORNITE DALLO SPECIALISTA ORTOPEDICO (ALLEGATA CONSULENZA). VISITA DI CONTROLLO A BREVE TERMINE PRESSO IL MEDICO DI LUOGO											1	

Our plan for an extra layer of de-identification

Exploratory analysis of the anonymized notes

Most False Positives are in headers that were not removed

This was an HEADER (we saw it earlier)

Veneto, Padova - Azienda Sanitaria Locale Città di Padova-PD - Direttore Dott. Mario Rossi
Presidi Ospedalieri ABCD – Corso Milano
Tel. 800 1234567 - +39 3216548709 (Segreteria)- Mail: example@for_the_presentation.com
Risultati validati dall'operatore tecnico sanitario nel pieno rispetto delle procedure e dei protocolli operativi predisposti dal Direttore della S.C. Lab. Analisi

Esame	Esito	U.M.	Val. Riferimento	Metodica
Sg-Emocromo				
WBC (Leucociti)	13.95 >	x 10^3 / ul	4.00 - 10.00	
RBC (Eritrociti)	5.24	x 10^6 / ul	4.50 - 6.00	
Monociti val.assoluto	0.96	x 10^3 / ul		

[...]

S-Lipasi	34	U/l	13 - 60	
S-Proteina C reattiva	2.3 >	mg/dl	0.1 - 0.5	
S-Troponina T	0.04	ng/mL	< 0.03	

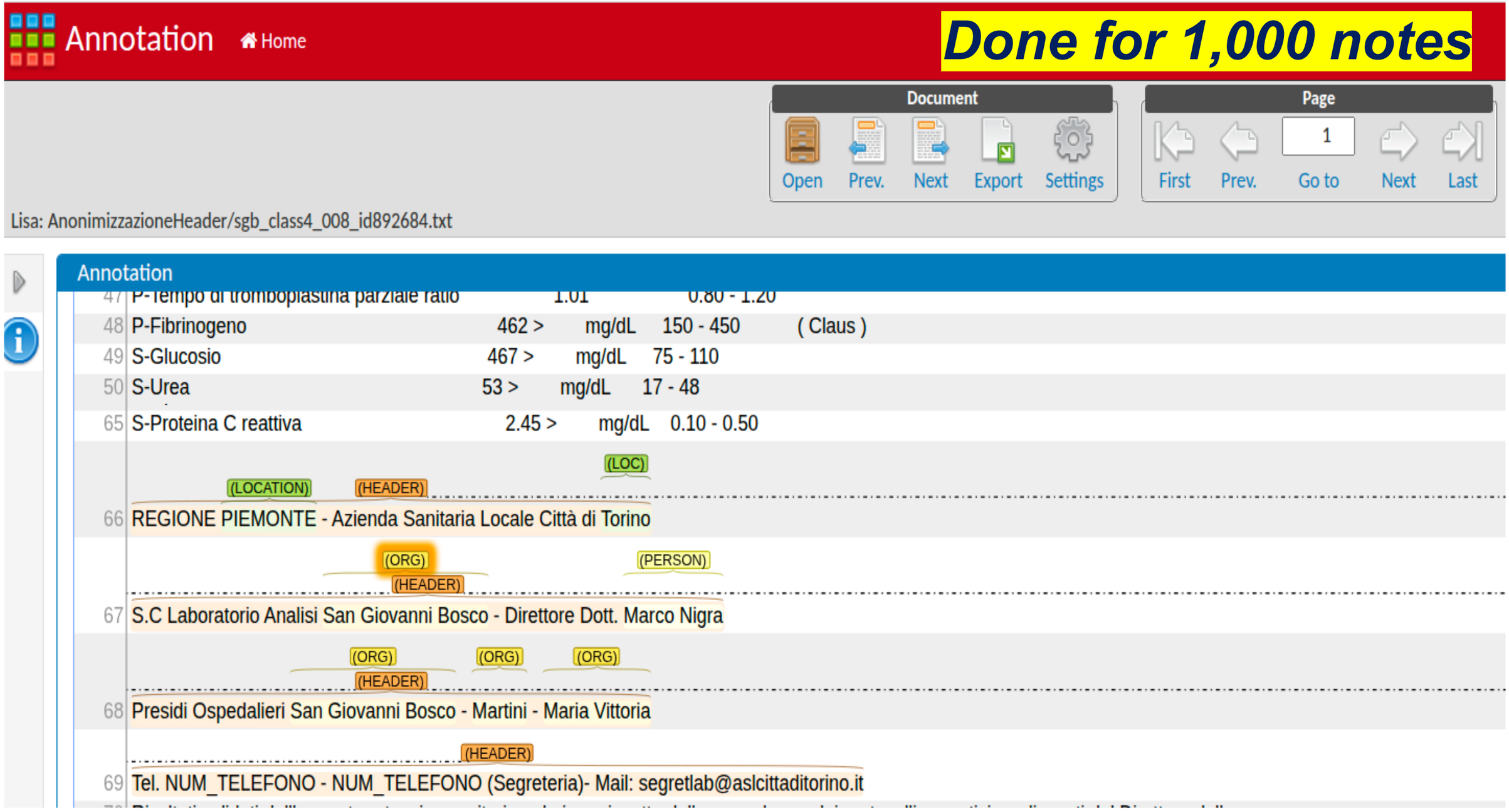
Troponina T eseguita su campione emolizzato, possibili interferenze analitiche. Ripetere prelievo.
Legenda Materiali: S = Siero P = Plasma Sg = Sangue U = Urine dU = Diuresi Oraria La= Liquidi Altri LCR = Liquor F = Feci

Our plan for an extra layer of de-identification

Exploratory analysis of the anonymized notes

Most False Positives are in headers that were not removed

Manual annotation: find missed headers and identifiers


Annotation
Home
Done for 1,000 notes
Document
Open
Prev.
Next
Export
Settings
Page
1
First
Prev.
Go to
Next
Last
Lisa: AnonimizzazioneHeader/sgb_class4_008_id892684.txt
Annotation
47 P-Tempo di tromboplastina parziale ratio 1.01 0.80 - 1.20
48 P-Fibrinogeno 462 > mg/dL 150 - 450 (Claus)
49 S-Glucosio 467 > mg/dL 75 - 110
50 S-Urea 53 > mg/dL 17 - 48
65 S-Proteina C reattiva 2.45 > mg/dL 0.10 - 0.50
(LOC)
(LOCATION)
(HEADER)
66 REGIONE PIEMONTE - Azienda Sanitaria Locale Città di Torino
(ORG)
(PERSON)
(HEADER)
67 S.C Laboratorio Analisi San Giovanni Bosco - Direttore Dott. Marco Nigra
(ORG)
(ORG)
(ORG)
(HEADER)
68 Presidi Ospedalieri San Giovanni Bosco - Martini - Maria Vittoria
(HEADER)
69 Tel. NUM_TELEFONO - NUM_TELEFONO (Segreteria)- Mail: segretlab@aslcittaditorino.it

Our plan for an extra layer of de-identification

Exploratory analysis of the anonymized notes

Create a system that fixes this issues

Most False Positives are in headers that were not removed

Manual annotation: find missed headers and identifiers

REMOVING THE HEADERS

	Recall	
CROSS		
sgb → vercelli	1.2%	*Learning to find the issues on one hospital, applying it to another one*
vercelli → sgb	29.6%	
PARITY		
sgb (n=126)	98.2%	*Learning to find the issues on one hospital, applying it to the same one*
vercelli (n=126)	90.6%	
POOLED		
tutti gli ospedali	98.8%	*Learning on both, applying it to both*

OUTSIDE THE HEADER

	Recall
CROSS	
sgb → vercelli	0.0%
vercelli → sgb	6.3%
PARITY	
sgb (n=126)	6.3%
vercelli (n=126)	0.0%
POOLED	
tutti gli ospedali	5.3%

Learning to find the issues on one hospital, applying it to another one

Learning to find the issues on one hospital, applying it to the same one

Learning on both, applying it to both

Current situation

AnonimAI performance: ~96.7% recall

Extra-layer performance: ~5% recall

How many identifiers we miss? 95% of the 3.3% ---> 2.8%

Current situation

This is what we can impact on

AnonimAI performance: ~96.7% recall

Extra-layer performance: ~5% recall

How many identifiers we miss? 95% of the 3.3% ---> 2.8%

And it might take a while
(on the 5k notes x 140 items was a week)

The «Positive Challenges» in Bernardo's presentation

Leveraging these results

Get other research groups involved in our problem (de-identification) and data



CLEF 2027 BUCHAREST

September 14-17, 2027

Bucharest, Romania

LISN
LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

eHealth Lab @ CLEF 2027

PLATEFORME DES DONNÉES DE SANTÉ
FRENCH HEALTH DATA HUB

alt-edic

Process

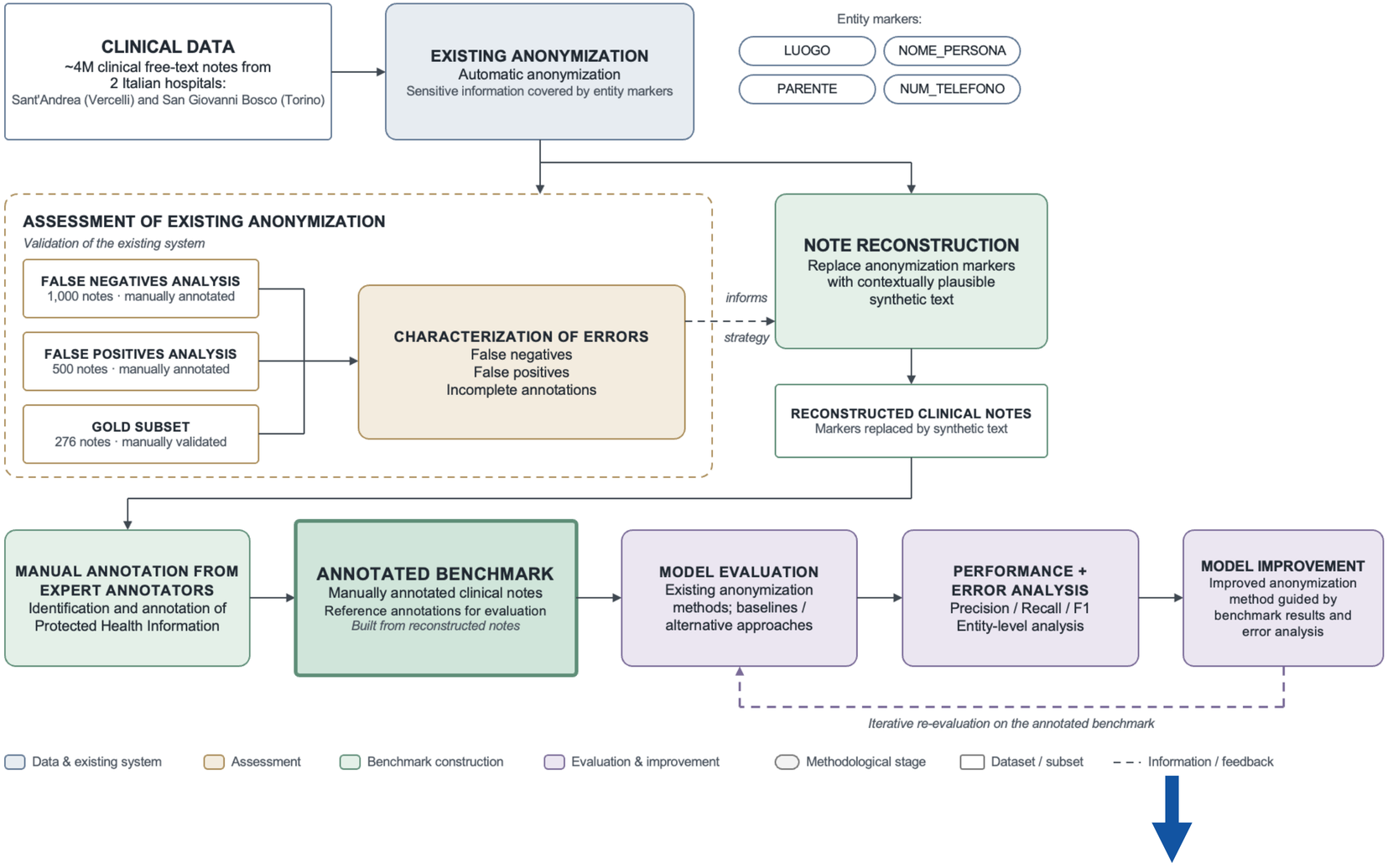
We cannot release the original notes with identifiable information

↓

Create «smart» surrogates for the de-identified notes → Ask humans to remove personal idetifiers (2,000 notes)

↓





eCREAM

de-identify

Thank you!



FONDAZIONE
BRUNO KESSLER

FONDAZIONE
BRUNO KESSLER