

Machine learning algorithms for Natural Language Processing in stroke discharge reports

ictusnet-sudoe.eu

Interreg



Sudoe

ICTUSnet 
European Regional Development Fund

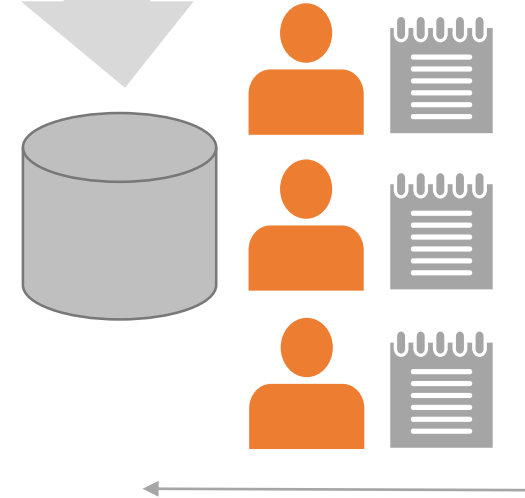
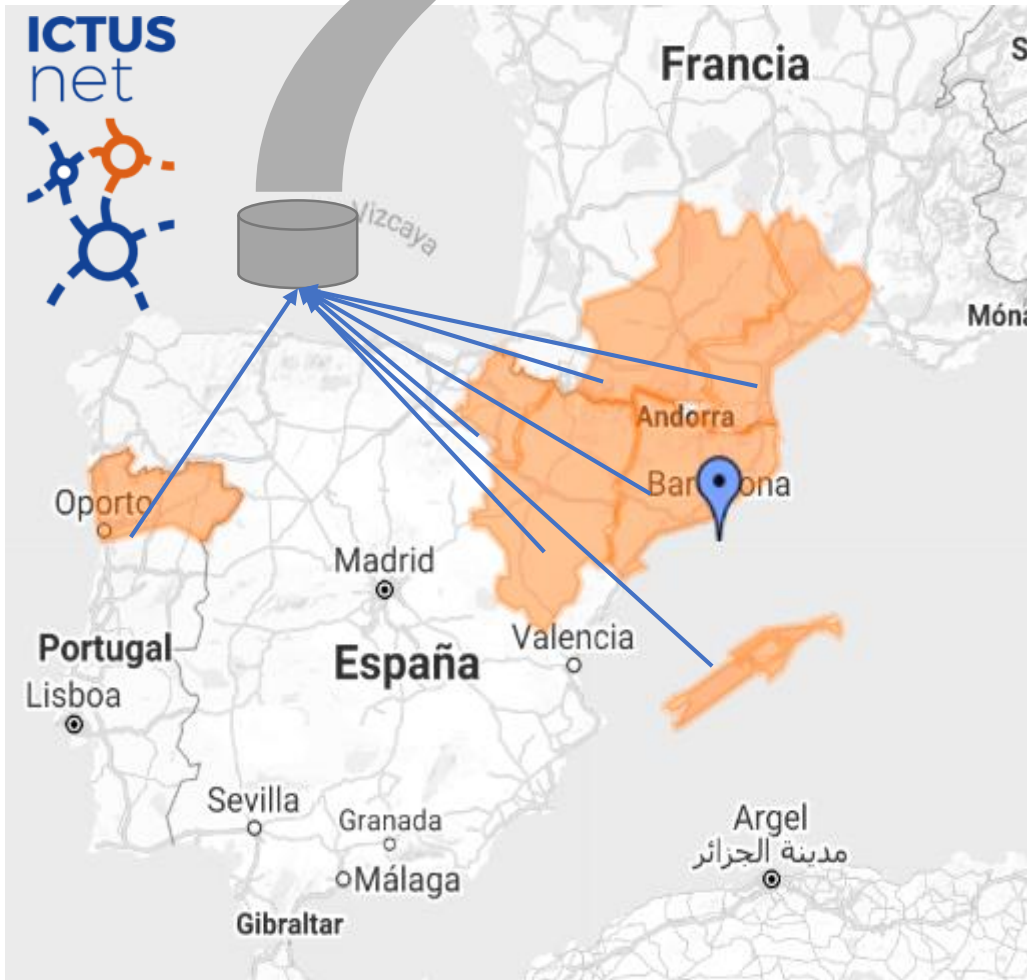
Marta Villegas
Aitor González, Siamak Barzegar,
Casimiro Carrino, Jordi Armengol, Asier Gutiérrez



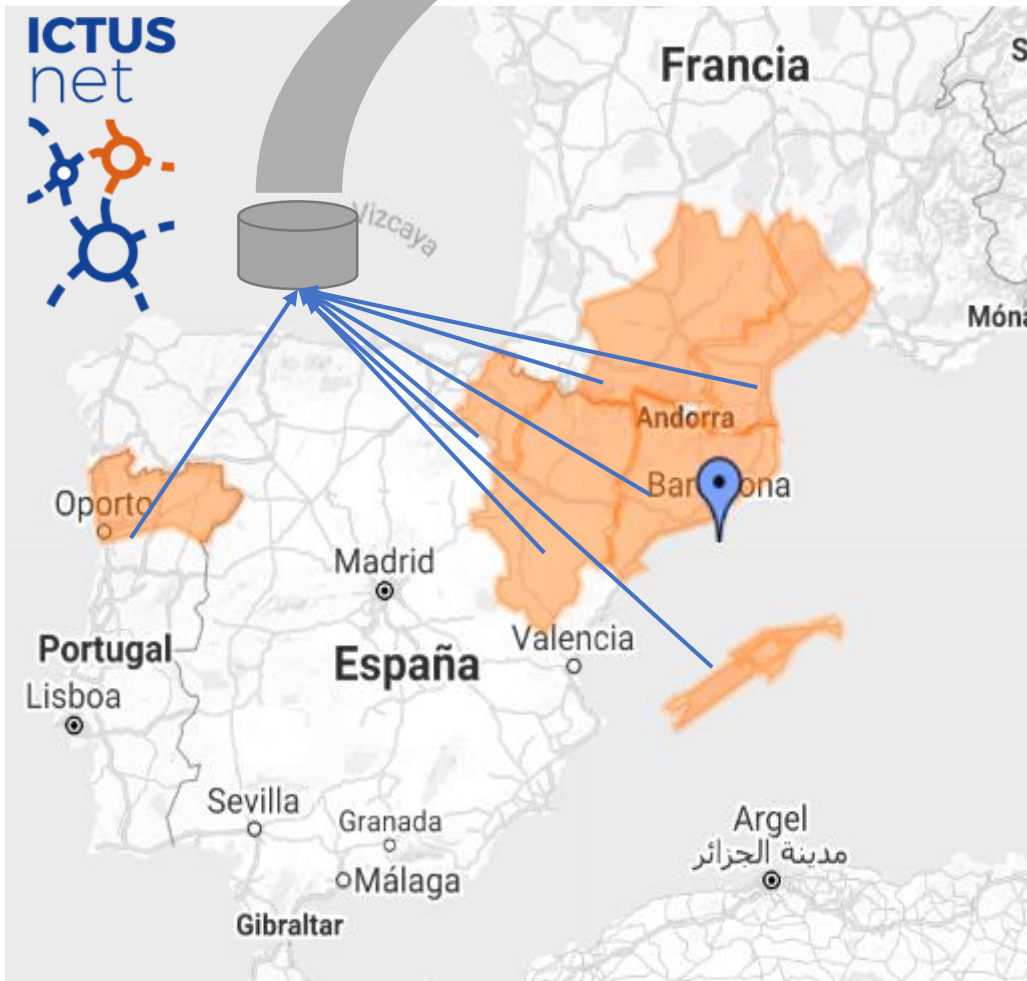
**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación

“INNOVATIVE USES OF REAL WORLD DATA ON STROKE”

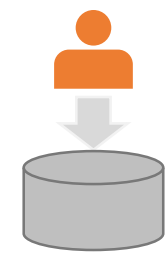
An information extraction service to support human experts



An information extraction service to support human experts

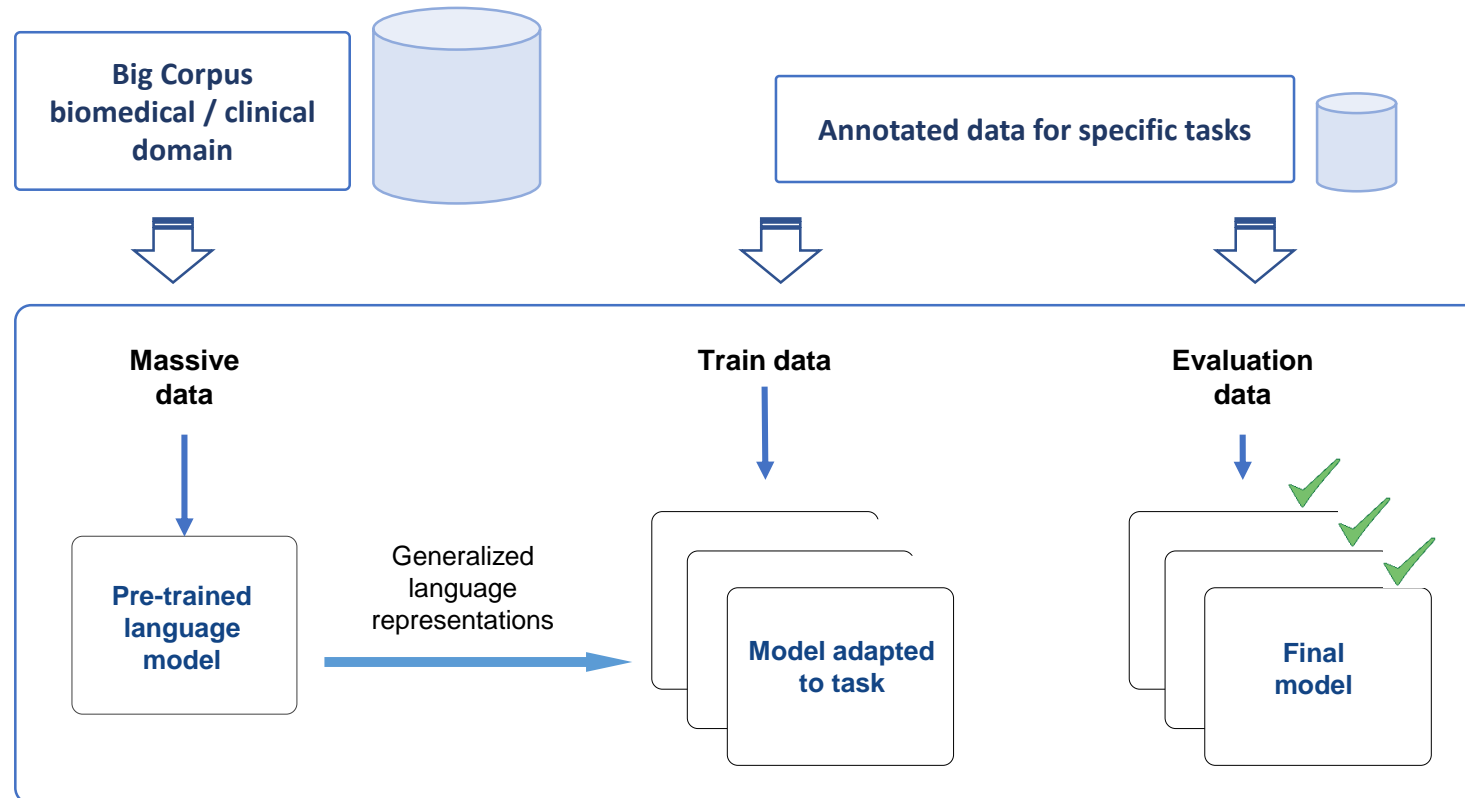


The screenshot displays the ICTUSnet web interface. On the left, there is a navigation menu with categories: 'Entrada y salida del paciente', 'Diagnóstico', 'Procedimientos I - trombolisis', 'Procedimientos II - trombectomía', and 'Tratamientos'. Below this is a section for 'Pruebas y escalas de valoración' with a 'TAC craneal' entry for '17/01/2017' and an 'ASPECTS' score of '10'. The main content area shows a detailed radiology report for 'Radiología / Radiología' with the following text: '*RADIOGRAFIA DE TORAX: ICT dentro de la normalidad. No se observan infiltrados ni condensaciones. *TC CRANEAL MULTIMODAL: - TC CRANEAL BASAL: No hay hemorragia intracraneal. Parénquima cerebral sin pérdida de la diferenciación cortico- subcortical. ASPECTS 10. - TC PERFUSION CEREBRAL: Territorio afectado: superficial de la ACMI. TTP: Retraso a nivel frontal, parietal, temporal superior e insular. CBF: Disminuido en territorio prácticamente superponible al de retraso del TTP. CBV: No se observan focos de caída del volumen cerebral. Porcentaje del área penumbra: >200%. - ANGIO-TC EXTRA E INTRACRANEAL: Leve ateromatosis que no condiciona estenosis en arco aórtico y salida/segmentos proximales de troncos supraaórticos. Ateromatosis calcificada en ambas bifurcaciones carotídeas, con estenosis de aprox. el 50% en el origen de la carótida interna izquierda. Se aprecia un STOP en el paso de contraste en una rama M1 distal/M2 proximal de la ACMI. El resto de principales arterias cerebrales intracraneales. *TC CRANEAL DUAL: Hiperdensidad lineal en cisura silviana izquierda así como otros focos de hiperdensidad a nivel temporal posterior y parietal izquierdo, que tras el postprocesado de las imágenes corresponde prácticamente en su totalidad a sangrado agudo. *TC CRANEAL CONTROL (17/1): Resolución parcial de hemorragia en cisura silviana. No otras alteraciones.'

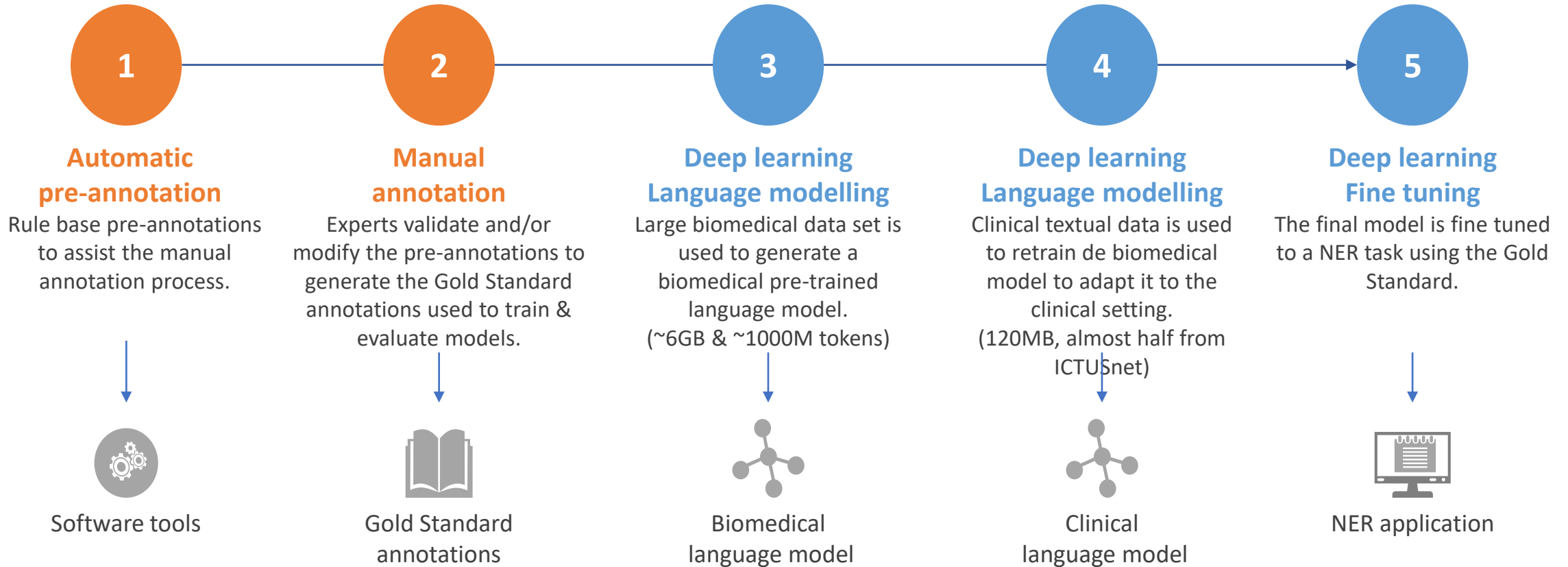


Methodology

1. We used deep learning techniques and deep neural networks to generate a **domain specific language model**.
2. Then, we adapt this model (**fine tune** it) to a specific task (Named Entity Recognition, NER)



Methodology



INDEX

1. Annotation variables & main challenges
2. (Pre)-annotation task
3. Language modelling & fine-tuning to NER
4. Evaluation
5. Summary & Conclusions
6. Examples & demo

VARIABLES and main challenges

ICTUSnet includes **51 different variables of interest** to be identified in discharge clinical reports.

These can be grouped into 6 groups:

- Section headers
- Main diagnosis & associated variables
- Procedures & associated temporal information
- Treatments
- Scales
- Temporal variables

```
/joint_files/all/321459759.utf8
Tratamiento_anticoagulante_hab
28 Fibrilación auricular en tratamiento con sintrom.
29 5.
30 Vertígen posicional poroxístico benigno en tratamiento médico Intervenciones quirúrgicas: cataratas de ambos ojos.
SECCION SITUACION FUNCIONAL
31 Situación basal: Independiente para las ABVD, continente biesfinteriana.
mRankin_previa
32 Barthel: 85, mRS0
SECCION TRATAMIENTO HABITUAL Tratamiento_anticoagulante_hab
33 Tratamiento habitual: Serc 8 mg 1c/24h, sintrom según pauta, pravastatina 40 mg 1c/24h, bisoprolol
SECCION PROCESO ACTUAL
35 ENFERMEDAD ACTUAL:
36 Paciente de 84 años el día 02/01/2017 a las 10:00am en la sala de espera de oftalmología inicia clínica de afasia motora, her
37 A su llegada a urgencias la paciente presenta afasia, hemiplejía derecha, babilinski derecho y desviación de la mirada hacia la
TAC craneal
38 Se le realiza un TAC craneal simple que descarta lesiones expansivas o sangrado agudo, se realiza posteriormente una Angio
Trombectomia_mecanica
penebra por lo que se envía al _SS_ para realizar embolectomía mecánica.
39 En el _SS_ se realiza microcateterismo selectivo de la arteria ocluida, logrando atravesar el trombo a las 13 horas y 15 minu
40 Se realiza extracción mecánica con sistema TREVO 6 X 25 y aspiración manual desde CAD sofía 5F obteniendo revascularizac
41 Se da por concluido el procedimiento sin complicaciones a confirmándose en la arteriografía un grado de reperusión grado 3
TAC craneal
42 Se realiza un TAC craneal posterior que muestra Hiperdensidad en región temporal anteromedial izquierda y en ganglios de l
43 Tras el procesado de las imágenes se evidencia que mayormente corresponde a contraste, identificando un tenue sangrado e
44 Hipodensidad con desdiferenciación corticosubcortical temporal posteromedial en relación con lesión isquémica establecida.
45 Se realiza traslado a HUMT y se ingresa en Neurología para completar estudio.
SECCION EXPLORACION FISICA
EXPLORACIÓN FÍSICA:
```

Section headers (need to normalize)

- Data coming from different providers with different structures and formats, need to normalize (use of Arquetipos suggested by Spanish Health Ministry as a mapping standard)).
- Some variables are 'context dependent' (they are only relevant provided they are in a specific section)
- We need *zoning* to discriminate variables by context

DIAGNÓSTICOS:

1. ICTUS TACI DE ACM IZQUIERDA
2. FIBRINÓLISIS IV + UROKINASA INTRAARTERIAL
3. HIPERTENSIÓN ARTERIAL
4. DISLIPEMIA
5. FORAMEN OVAL PERMEABLE

Paciente varón de 66 años, sin alergias conocidas.

*Hábitos tóxicos: fumador de 15 cig/día.

*Sociofuncional: independiente para las ABVD, vive con su esposa. Rankin 0

PROCEDIMIENTOS:

TC craneal

Angio TC craneal

Fibrinólisis IV con alteplasa

Arteriografía con administración de urokinasa local.

Monitorización en UCI

Monitorización en Unidad de Ictus

ANTECEDENTES:

- Hipertensión arterial en tratamiento con 2 fármacos.

- Dislipemia en tratamiento dietético.

- Lumbociatálgia izquierda L5. En seguimiento por Reumatología y Unidad del dolor.

- Espondiloartrosis cervical C5-C7 con calcificación de tejidos blandos (miositis osificante).

- Neuropatía bilateral cubital en tratamiento con Gabapentina.

MEDICACIÓN HABITUAL:

Enalapril 10mg/12h, Hidroclorotiazida 12.5mg/24h, Gabapentina 150mg/12h, Domperidona 10mg/12h

Zaldiar 50mg/12h ÍNDEX BARTHEL PREVI : 100

ENFERMEDAD ACTUAL:

Paciente traído el 29/11 a las 23:18 como código ICTUS extrahospitalario. Presenta cuadrante derecho, afasia y desviación de la comisura bucal de inicio a las 21:45h.

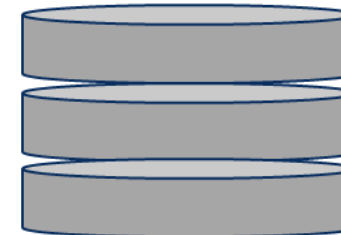
Set of variables

V1

V2

V3

V4



Section headers (and structural heterogeneity)

SECCION MOTIVO DE INGRESO

MOTIU D'INGRÉS

Paciente que ingresa por ictus ACM izq y ACA izq cardioembólicas.

SECCION ANTECEDENTES

ANTECEDENTS

No alergias conocidas No hábitos tóxicos FRCV:

-No HTA, no DLP, no DM

-Cardiopatías: FA anticoagulada con Sintrom (control en privada Dr. _NAME_ centro.), valvulopatía mitral no tributaria

-Niega otros antecedentes de interés.

SECCION ANTECEDENTES QUIRURGICOS

IQ:

osteosíntesis húmero proximal izq. apendice y colecistectomía. mrankin 0.

Vive sola.

Deamulación autónoma.

Independiente para las ABVD.

SECCION TRATAMIENTO HABITUAL

MEDICACIÓ HABITUAL

CO-VALS FORTE 160MG/25MG VALSARTAN+DIURETIC 1 x 24 h.

Indefinida EMCONCOR COR 2,5MG BISOPROLOL, FUMARAT 1 x 24 h.

Indefinida SINTROM 4MG ACENOCUMAROL 1 x 24 h.

Indefinida

SECCION PROCESO ACTUAL

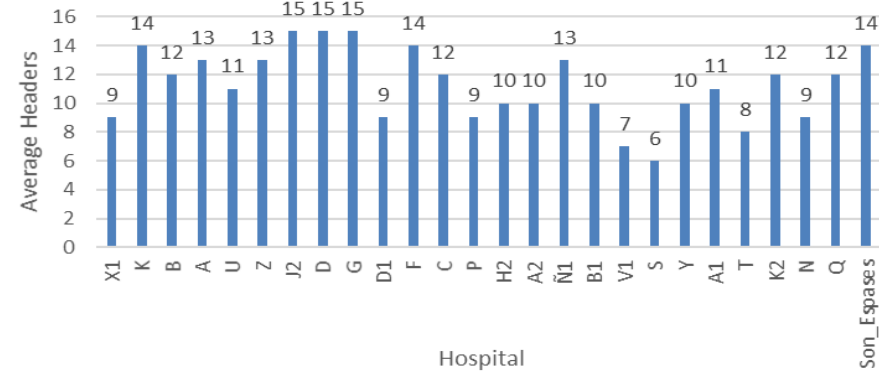
PROCÉS ACTUAL

Fecha llegada hospital

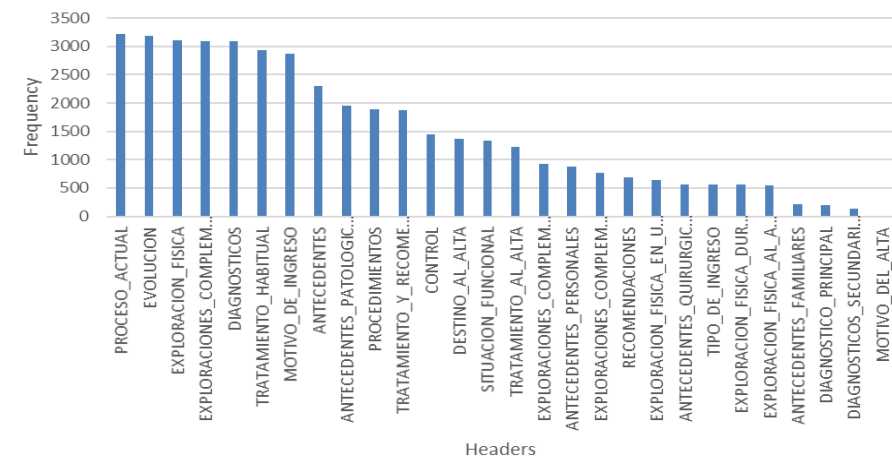
Paciente de 82 años que el día 11/01 es encontrada por familiares a las 19.40h en domicilio acostada en quien objetiva RACE 8 y activa código ictus desde domicilio a las 20.15h.

La paciente estaba vestida con ropa de cama (ella siempre va a la cafetería a desayunar y hoy no la han visto, por lo q

Average headers per file



Frequency of headers in the whole dataset



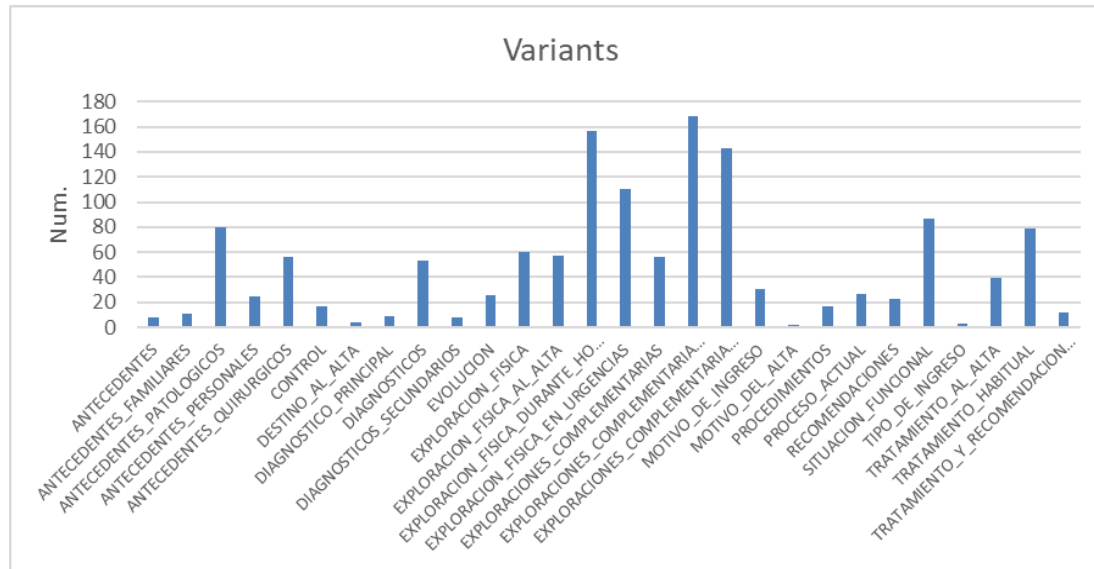
Section headers (and naming conventions)



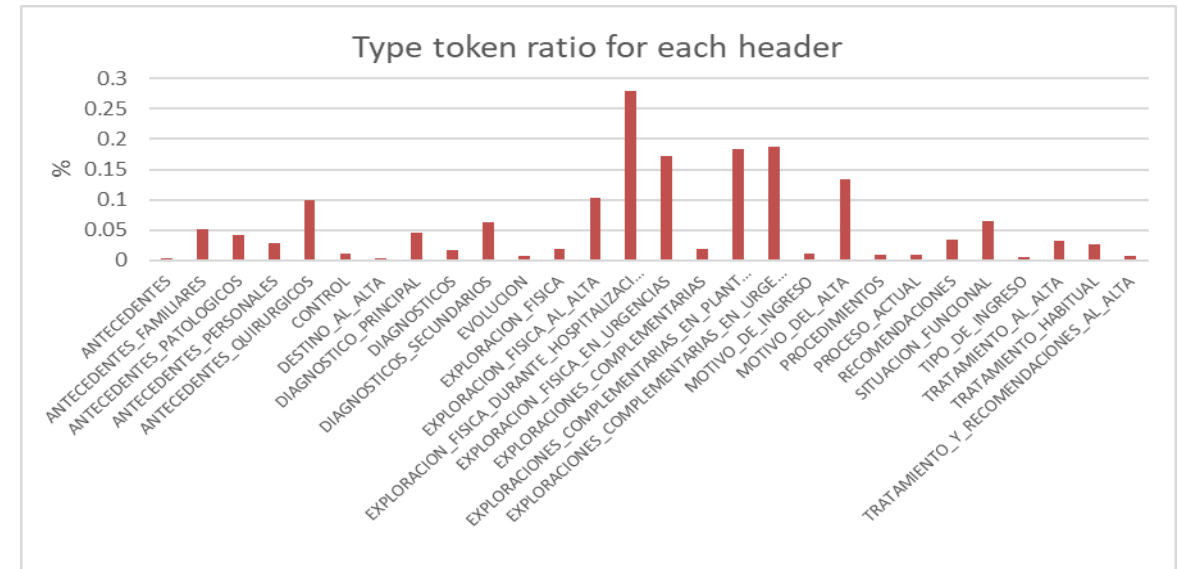
| | | |
|---------------------------------|-----------------------------|-------------------------------|
| ANTECEDENTES_PATOLOGICOS | | |
| A. patològics | ANTECEDENTES PATOLÒGICOS | ANTECEDETES PATOLÓGICOS |
| Anecedentes mèdics | Antecedentes patològics | ANTECEDETNES MÉDICOS |
| ANTECEDENTES PATOLÓGICOS | ANTECEDENTES PATOLOGICS | Antecedetnes Patològics |
| ANTECEDENTES PATOLÓGICOS | ANTECEDENTES PATOLÓGICS | ANTECEDNETES PATOLÓGICOS |
| ANTECEDENTES MÉDICOS | ANTECEDENTES PATOLÒGICS | Antecednetes Patològics |
| ANTECEDENETS PATOLOGICOS | ANTECEDENTS PATOLOGICS | ANTECEDENTES PATOLÓGICOS |
| Antecedente Patologicos | Antecedents patològics | Antecedentes patològics |
| ANTECEDENTE PATOLOGICOS | ANTECEDENTS PATOLÒGICS | ANTECEDETES PATOLÓGICOS |
| ANTECEDENTE PATOLÓGICOS | ANTECEDENTS MÉDICOS | ANTECENTES PATOLÓGICOS |
| ANTECEDENTES PATOLÓGICOS | ANTECEDENTS MÉDICOS | ANTECENTS PATOLÒGICS |
| ANTECEDENTES PATOLÒGICOS | Antecedents mèdics | Antedentes Patològics |
| Antecedentes Medicos | ANTECEDENTS PATOLÒGIC | ANTEEDENTS PATOLÒGICS |
| ANTECEDENTES MEDICOS | ANTECEDENTS PATOLOGIC S | APatològics |
| Antecedentes mèdics | ANTECEDENTS PATOLÒGIC S | HISTORIAL MÈDIC |
| ANTECEDENTES MÉDICOS | ANTECEDENTS PATOLOGICOS | Historial mèdic |
| Antecedentes Médicos | ANTECEDENTS PATOLÓGICOS | HISTORIAL MEDICO |
| Antecedentes mèdics | Antecedents Patològics | HISTORIAL MÉDICO |
| ANTECEDENTES PAOTLÓGICOS | ANTECEDENTS PATOLÒGICOS | Historial mèdic |
| ANTECEDENTES PATOLIOGICOS | ANTECEDENTS PATOLOGICS | Malalties prèvies |
| Antecedentes patològics | Antecedents patològics | Otros antecedentes patològics |
| ANTECEDENTES PATOLOGICOS | ANTECEDENTS PATOLÓGICS | PATOLÓGICAS |
| Antecedentes Patologicos | ANTECEDENTS PATOLÒGICS | PATOLOGICOS |
| Antecedentes patologicos | Antecedents patològics | PATOLÓGICOS |
| Antecedentes patològics | Antecedents Patològics | Patològics |
| ANTECEDENTES PATOLÓGICOS | ANTECEDENTS PATÒLOGICS | PATOLOGICS |
| Antecedentes Patològics | Antecedents patològics FRCV | PATOLÒGICS |
| ANtecedentes Patològics | Antecedents patoògics | |

The 80 variants for ANTECEDENTES PATOLOGICOS

Section headers (and naming conventions)



Number of variants per header (i.e. different ways to mention a header)



TTR is the total number of unique headers (types) divided by the total number of headers (tokens) in the documents

Section headers (and format heterogeneity)

SECCION MOTIVO DE INGRESO

MOTIU D'INGRÉS

Catalan / uppercase / no special markup / full line

SECCION TRATAMIENTO HABITUAL

* MEDICACIÓN HABITUAL: insulina asp

Spanish / uppercase / special markup / inserted in paragraph

SECCION PROCESO ACTUAL

Proceso actual / Procés actual

Bilingual / lowercase / no special markup / full line

SECCION SITUACION FUNCIONAL

- Situación sociofuncional: Vive solo, no tiene hijos.
Trabajó de agente comercial. Independiente para todas las ABVD.

Spanish / lowercase / special markup / inserted in paragraph

SECCION PROCEDIMIENTOS

Proc.:

ANALÍTICA SANGUÍNEA () RADIOGRAFÍA DE TÓRAX () ECG.

? / abbreviation / lowercase / no markup / full line

← Upper case!!

Main diagnosis and associated variables

This includes three main diagnoses: *ictus isquémico*, *ataque isquémico transitorio* and *hemorragia cerebral* and their associated attributes:

- affected vessel,
- localization,
- lateralization and
- etiology.

Context dependent variables
(only relevant if in DIAGNOSTICOS section)

SECCION DIAGNOSTICOS

DIAGNÒSTICS

SUG_Ictus_isquemico

SUG_Arteria_afectada

SUG_Lateralizacion

SUG_Localizacion

SUG_Etiologia

- Ictus isquèmic de territori de ACM esquerra (big lacunar) de etiologia indeterminada
 -Hipertensió arterial essencial
 -Diabetis Mellitus tipus 2
 -Dislipèmia
 -Glaucoma

Procedures & associated temporal information

Procedures

Trombolisis_intravenosa
 Trombectomia_mecanica
 Trombolisis_intraarterial
 Test_de_disfagia
 Tac_craneal

Associated temporal information:

Fecha / hora Tc craneal inicial
 Fecha trombólisis iv
 Hora inicio primer bolus de la trombólisis rtPA
 Fecha trombectomía mecánica
 Hora punción arterial para la trombectomía mecánica
 Fecha / hora primera serie trombectomía mecánica
 Hora trombólisis intraarterial
 Fecha / hora recanalización
 Fecha / hora finalización trombectomía
 Fecha / hora trombólisis intraarterial

Hora primer bolus trombolisis rtPA

Trombolisis intravenosa

La pacient compleix criteris per rebre tractament fibrinolític amb rtPA que s'administra a les 11'37h sense incidències.

```
T30 Trombolisis_intravenosa 2568 2600 tractament fibrinolític amb rtPA
T14 Hora_primer_bolus_trombolisis_rtPA 2568 2631 tractament fibrinolític amb rtPA que s'administra a les 11'37h
#3 AnnotatorNotes T14 11:37
```

TAC_craneal

Fecha_TAC


TAC_craneal

-TC craneal de control a las 24h (22/3/17): En valoración comparativa con estudio TC previo no identfico

Treatments

For **treatments**, the objective is to find **anticoagulants** and **antiaggregants** and to classify them as “**pre admission medication**” or “**discharge medication**”.

This task is essentially a NER task that includes a classification part (pre-admission vs discharge). This classification mostly depends on the context of the mention (i.e. the section in which the medication is listed).



SECCION TRATAMIENTO AL ALTA

Plan terapéutico / Pla terapèutic

- Metoclopramida 10 mg/8h.
- Esomeprazol 20 mg/24h.

Tratamiento antiagregante alta

- Ácido acetilsalicílico 100 mg/24h.
- atorvastatina 40 mg/24h.
- Paracetamol 1 g/8h.
- Bisoprolol 5 mg/24h.
- Captopril 25 mg/8h.
- Alopurinol 100 mg/24h.

Scales



The relevant **scales** to be annotated include:

ASPECTS

mRankin_previa, mRankin_alta
NIHSS_previa, NIHSS_alta

The main challenge in this case is to find the numerical value of the rating scale (often this comes in a complex format) and to distinguish between *pre-admission vs discharge* categories.

Note that in the vast majority of cases, this distinction is not explicitly expressed in the reports. Again, these are context dependent variables.

NIHSS_previa
Escala NIHSS (0-2-0-0-2)+(1-0-0-0-0)+(0-1-2-2-0)= 10.

SECCION ANTECEDENTES QUIRURGICOS

IQ:

mRankin_previa
osteosíntesis húmero proximal izq. apendice y colecistectomía. mrankin 0.

SECCION SITUACION FUNCIONAL

Estado basal: viu amb el marit, autonoma per les AVD.

mRankin_previa
Rankin previo 2

VARIABLES and main challenges

- Multiple providers
 - Heterogeneous data with different
 - formats, structure, headers, naming conventions, ...
- Nature of Language
 - Acronyms and abbreviations.
 - Ambiguity.
 - Negation
- User generated content
 - Ungrammatical/telegraphic sentences.
 - Text in different languages.
 - Text written by different personnel.
 - Typos.

Mujer de 84 años sin **ACM**. Niega hábitos tóxicos. Parcialmente dependiente **ABVD**. Vive en residencia. Antecedentes de **HTA**, **DLF** y **FA** antiagregada. Ictus **POCI ACP** izquierda en 2008, etiología cardioembólica.

Mujer de 84 años sin **alergias medicamentosas conocidas**. Niega hábitos tóxicos. Parcialmente dependiente **adriamycin bleomycin vinblastine and dacarbazine**. Vive en residencia. Antecedentes de **hipertensión arterial, depresión a largo plazo** y **fibrilación auricular** antiagregada. **Ictus circulación posterior, arteria cerebral posterior** izquierda en 2008, etiología cardioembólica.

TC craneal urgente multimodal: No es visualitzen signes hemorràgics aguts intra ni extraxials.

TC de control (24h post tratamiento fibrinolítico): Infart fronto-parieto-temporal dret establert, en territori d'ACM dreta, sense signes de transformació hemorràgica

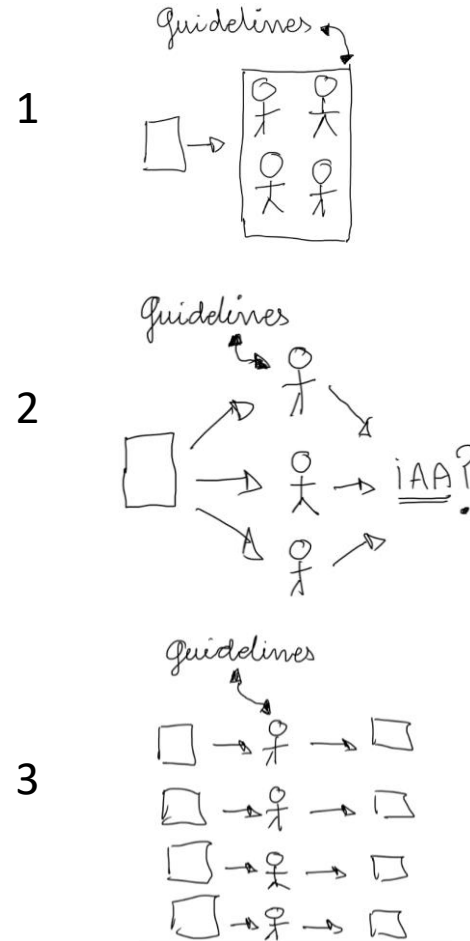
Fàcil masticació. 1º Plato sopa o pure.

Durante el ingreso en la unidad de ictus, presenta un empeoramiento neurológico, por lo que se realiza un TC craneal de control que evidencia una nueva lesión isquémica en ACP izquierda. También presenta una fibrilación auricular de debut, por lo que se decide iniciar tratamiento anticoagulante. Avui control INR (s'enviarà resultat a IAS).

AngioTC craneal (6.12.2017): Poligono de Willis normoconfigurado, sin imagenes de adiccion ni defectos de replección. No se observan imagenes sugestivas de malformación vascular. Segmento V4 esquerre hipoplàsico.

Annotation task

Annotation task is difficult and needs good annotation guidelines, good protocols and a friendly annotation tool.



[/joint_files/all/321459759.utf8](#)

28 Fibrilación auricular en tratamiento con Tratamiento_anticoagulante_hab sintrom.

29 5.

30 Vertígen posicional poroxístico benigno en tratamiento médico Intervenciones quirúrgicas: cataratas de ambos ojos.

SECCION SITUACION FUNCIONAL

31 Situación basal: Indenpendiente para las ABVD, continente biesfinteriana.

mRankin_previa

32 Barthel: 85, mRS0

SECCION TRATAMIENTO HABITUAL Tratamiento_anticoagulante_hab

33 Tratamiento habitual: Serc 8 mg 1c/24h, sintrom según pauta, pravastati

SECCION PROCESO ACTUAL

35 ENFERMEDAD ACTUAL:

36 Paciente de 84 años el día 02/01/2017 a las 10:00am en la sala de espera de oftalmología inicia

37 A su llegada a urgencias la paciente presenta afasia, hemiplejia derecha, babinski derecho y des

TAC_craneal

38 Se le realiza un TAC craneal simple que descarta lesiones expansivas o sangrado agudo, se reali

Trombectomia_mecanica

penumbra por lo que se envía al _SS_ para realizar embolectomía mecánica.

39 En el _SS_ se realiza microcateterismo selectivo de la arteria ocluida, logrando atravesar el trom

40 Se realiza extracción mecánica con sistema TREVO 6 X 25 y aspiración manual desde CAD sofía 5

41 Se da por concluido el procedimiento sin complicaciones a confirmándose en la arteriografía un c

TAC_craneal

42 Se realiza un TAC craneal posterior que muestra Hiperdensidad en región temporal anteromedial

43 Tras el procesado de las imágenes se evidencia que mayormente corresponde a contraste, identi

44 Hipodensidad con desdiferenciación corticosubcortical temporal posteromedial en relación con le

45 Se realiza traslado a HUMT y se ingresa en Neurología para completar estudio.

SECCION EXPLORACION FISICA

i EXPLORACIÓN FÍSICA:

New Annotation

Text

embolectomia mecánica

Search

Google, Wikipedia

entity type

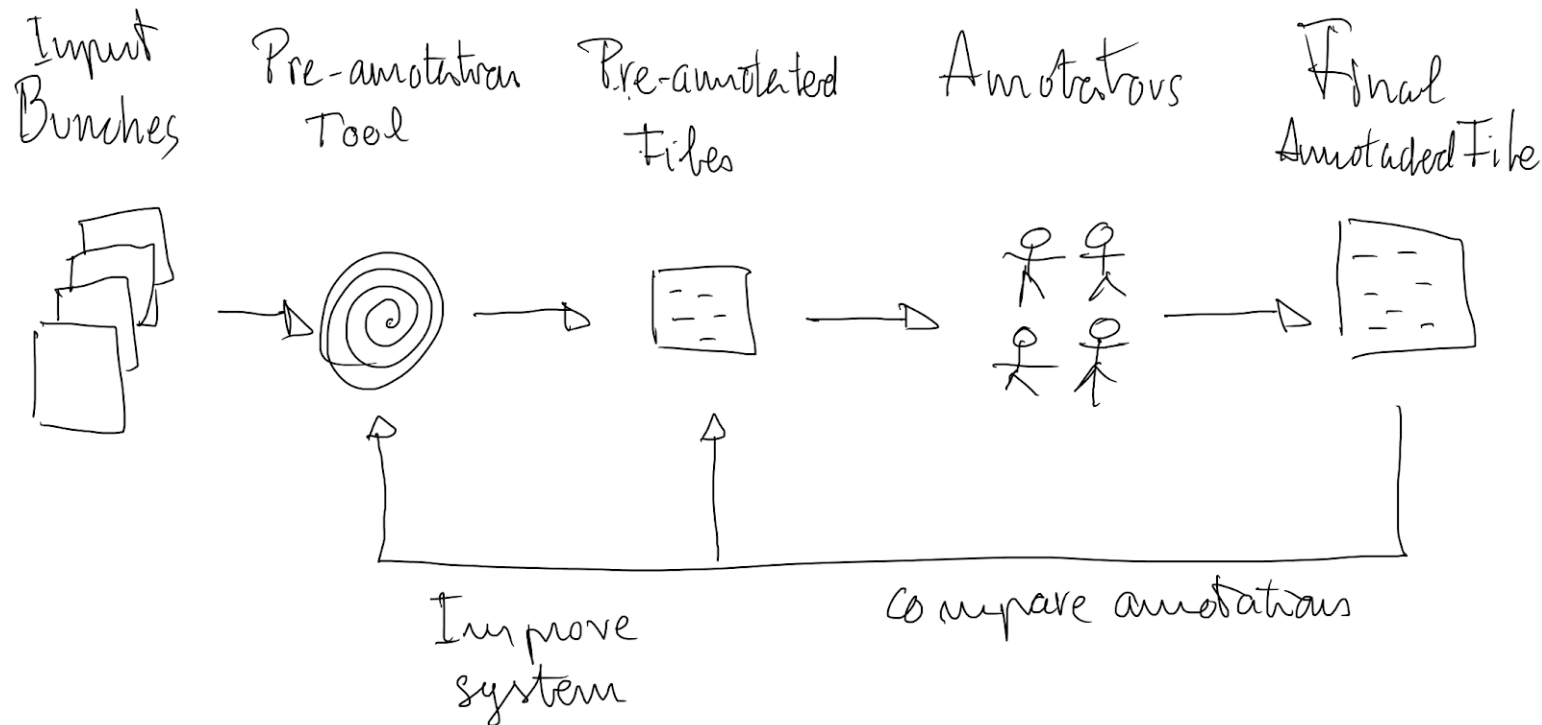
- DIAGNOSTICOS
- PROCEDIMIENTOS
 - Trombolisis_intravenosa
 - Trombectomia_mecanica**
 - Trombolisis_intraarterial
 - TAC_craneal
 - Test_de_distagla
 - Recanalizacion
 - Puerta_aguja
- TRATAMIENTOS
- ESCALAS
- PROG
- HORA
- TELEPRO
- SECCIONES

Notes

OK Cancel

(Pre-)annotation

To ease the task, an **automatic pre-annotation system** was developed in an **iterative way**, so that the process was split into different steps, each consisting of ~100 discharge reports. At each new bunch of annotated files, the system is evaluated against the human annotations and modified to improve its performance



(Pre-)annotation (evaluation)

| Num examples | True positives | False positives | False negatives | Accuracy | Precision | Recall | F1 |
|--------------|----------------|-----------------|-----------------|----------|-----------|--------|-------|
| 5710 | 5012 | 894 | 698 | 0.759 | 0.849 | 0.878 | 0.863 |

Global average results using the Test set

Language Modelling

3 We generated a **RoBERTa-base model** with 12 layers/heads and 768 hidden layer sizes for a total number of 126M parameters.

- We kept the original Roberta hyperparameter configuration and trained with a **masked language model** objective.
- The model was trained for 48 hours using 16 NVIDIA V100 GPUs of 16GB DDRAM.
- After training, we selected as the best model the checkpoint that achieved the lowest perplexity.

4 Then, we adapted the model to the clinical domain by **overtraining it** with 120MG of clinical textual data (including nearly 34MB of ICTUSnet data provided by AQUAS, Son Espases and IACS).

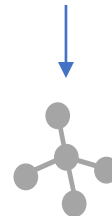
- We continued the training process for 48h more using the learning rate value reached by the best checkpoint trained on the biomedical corpora.

5 Finally, we **fine-tuned** our pre-trained models for NER task using the ICTUSnet Gold Standard dataset.

- The gold standard was split into train, dev and test sets with standard proportions: 80% for training (656 documents), 10% for valid (83 documents), 10% for test (83 documents).
- We fine-tuned for 10 epochs and selected the best epoch validating on the dev set.



**Deep learning
Language modelling**
Large biomedical data set is used to generate a biomedical pre-trained language model.
(~6GB & ~1000M tokens)



Biomedical
language model



**Deep learning
Language modelling**
Clinical textual data is used to retrain de biomedical model to adapt it to the clinical setting.
(120MB, almost half from ICTUSnet)



Clinical
language model



**Deep learning
Fine tuning**
The final model is fine tuned to a NER task using the Gold Standard.



NER application

Language Modelling

We collected a big **biomedical corpora** gathering from a variety of medical resources, namely scientific literature, clinical cases and crawled data.

We cleaned each corpus independently using a **cleaning pipeline** with customized operations designed to read data in different formats, split into sentences, detect the language, remove noisy and bad-formed sentences, ...

Finally we **deduplicate** and eventually output the data with their original document boundaries.

| <u>Corpus name</u> | <u>Text Size (GB)</u> | <u>Final size (GB)</u> | <u>Raw tokens</u> | <u>Cleaned tokens</u> | <u>Num. sentences</u> |
|---------------------------|-----------------------|------------------------|-----------------------|-----------------------|-----------------------|
| Clinical cases cardiology | 0.0035 | 0,001 | 149,904.00 | 147,790.00 | 9,970.00 |
| Clinical cases radiology | 0.0066 | 0,001 | 177,366.00 | 170,997.00 | 9,948.00 |
| libros_casos_clinicos | 0.0083 | 0,007 | 1,137,555.00 | 1,024,797.00 | 68,833.00 |
| Clinical cases COVID | 0.0084 | 0,001 | 82,201.00 | 82,091.00 | 3,896.00 |
| EMEA corpus | 0.087 | 0,034 | 13,797,362.00 | 5,377,448.00 | 284,575.00 |
| Patents | 0.087 | 0,084 | 14,022,520.00 | 13,463,387.00 | 253,924.00 |
| wikipedia_life_sciences | 0.172 | 0,088 | 18,771,176.00 | 13,890,501.00 | 832,027.00 |
| barr2_background | 0.188 | 0,159 | 28,868,022.00 | 24,516,442.00 | 1,029,600.00 |
| Pubmed | 0.211 | 0,013 | 1,957,479.00 | 1,858,966.00 | 103,674.00 |
| REEC (casos clinicos) | 0.823 | 0,028 | 4,581,755.00 | 4,283,453.00 | 220,726.00 |
| mespen_medline | 1.2 | 0,38 | 6,864,901.00 | 4,166,077.00 | 322,619.00 |
| pdfs_general | 3.3 | | 09,124,996.00 | 7,146,139.00 | 5,252,481.00 |
| Scielo | 3.891 | 0,631 | 61,837,972.00 | 60,007,289.00 | 2,668,231.00 |
| Medical crawler | 606 | 4,5 | ? | 746,368,185.00 | 32,766,976.00 |
| TOTAL | 615.9858 | 5,927 | 261,373,209.00 | 972,503,562.00 | 43,827,480.00 |

Language Modelling

We used the resulting corpus to train a **RoBERTa-base model** with 12 layers/heads and 768 hidden layer sizes for a total number of 126M parameters.

- We kept the original Roberta hyperparameter configuration and trained with a **masked language model** objective.
- The model was trained for 48 hours using 16 NVIDIA V100 GPUs of 16GB DDRAM.
- After training, we selected as the best model the checkpoint that achieved the lowest perplexity.

Then, we adapted the model to the clinical domain by **overtraining it** with 120MG of clinical textual data (including nearly 1GB of ICTUSnet data provided by AQuAS, Son Espases and IACS).

- We continued the training process for 48h more.

Finally, we **fine-tuned** our pre-trained models for NER task using the ICTUSnet Gold Standard dataset.

- The gold standard was split into train, dev and test sets with standard proportions: 80% for training (656 documents), 10% for valid (83 documents), 10% for test (83 documents).
- We fine-tuned for 10 epochs and selected the best epoch validating on the dev set.

We used both, the Biomedical model and the Cincial model to generate and compare the predictions

Evaluation



Annotations are converted from **BRAT** standoff format to **BIO/IOB** (beginning, inside, outside) format.

- the prefix "B" in front of a Tag indicates the beginning of a chunk,
- an "I" indicates that we are still inside that chunk.
- the "O" tag is used to indicate that a token does not correspond to any of the entities to be tagged.

Given the high number of tokens with to the class "O", we do not consider them when predictions and GS we have an O label

this avoids raising the result due to the fact that O is the majority class, i.e., that the vast majority of tokens do not belong to any of the entities.

In this example, only the grey rows are evaluated. (in red the wrong ones and in Green the good ones).

Once the lines with double O are removed, we evaluate the model using **accuracy**, **precision**, **recall** and **F1**

Accuracy: number of correct predictions / total number of predictions

Precision: true positives / predicted positives (how many selected items are relevant)

Recall: true positives / actual positives (how many relevant items are selected)

F1: (balance between precision and recall)

| token | tag |
|---------------|------------------|
| Vive | O |
| con | O |
| su | O |
| esposa | O |
| , | O |
| independiente | O |
| para | O |
| ABVD | B-Antecedente |
| , | O |
| mRs | B-mRankin_previa |
| 0 | I-mRankin_previa |
| . | O |

Evaluation



Annotations are converted from **BRAT** standoff format to **BIO/IOB** (beginning, inside, outside) format.

- the prefix "B" in front of a Tag indicates the beginning of a chunk,
- an "I" indicates that we are still inside that chunk.
- the "O" tag is used to indicate that a token does not correspond to any of the entities to be tagged.

Given the high number of tokens with to the class "O", we do not consider them when predictions and GS we have an O label

this avoids raising the result due to the fact that O is the majority class, i.e., that the vast majority of tokens do not belong to any of the entities.

In this example, only the grey rows are evaluated. (in red the wrong ones and in Green the good ones).

Once the lines with double O are removed, we evaluate the model using **accuracy**, **precision**, **recall** and **F1**

Accuracy: number of correct predictions / total number of predictions

Precision: true positives / predicted positives (how many selected items are relevant)

Recall: true positives / actual positives (how many relevant items are selected)

F1: (balance between precision and recall)

| tken | GS | prediction |
|---------------|------------------|------------------|
| Vive | O | O |
| con | O | O |
| su | O | O |
| esposa | O | O |
| , | O | O |
| independiente | O | O |
| para | O | O |
| ABVD | O | B-Antecedente |
| , | O | O |
| mRs | B-mRankin_previa | B-mRankin_previa |
| 0 | I-mRankin_previa | I-mRankin_previa |
| . | O | O |

Evaluation

For time **variables** the evaluation is different. In this case, we have three components:

1. The textual evidence in text containing a time variable (text span)
2. The tag assigned (label)
3. The normalized time expression

When evaluating, we require that the predictions and the Gold standard

- for (2) and (3) above are equal, and
- there is some overlapping in (1) (textual evidences are long and 'irrelevant')

Examples:

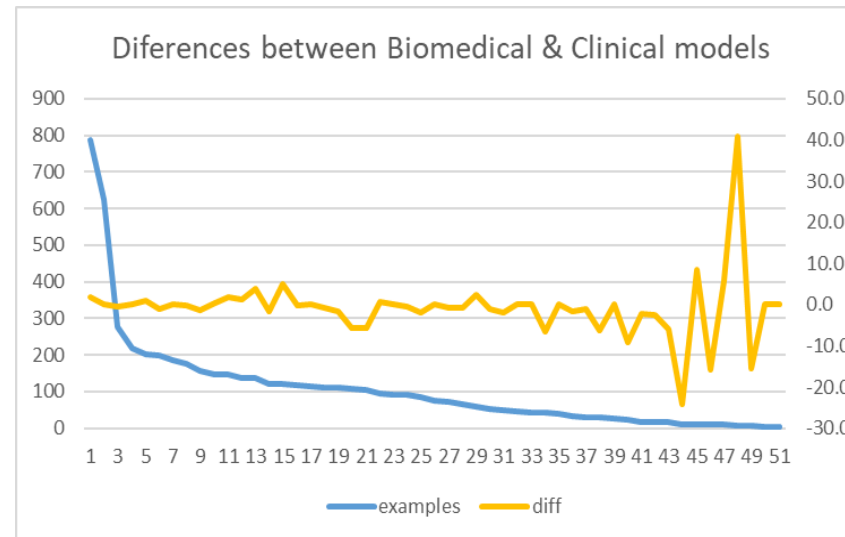
The diagram shows two examples of time variable mentions. In the first example, the text is "I (23/02/17; 13:03h) :". The time span "13:03h" is highlighted with a blue box labeled "Hora_TAC". To the right, a grey box contains the text "Hora_primer_bolus_trombolisis_rtPA" with a green box labeled "rTPA a las 13.33 h" underneath it. In the second example, the text is "(23/02/17; 13:03h) :". The time span "13:03h" is highlighted with a blue box labeled "Hora_TAC". To the right, a grey box contains the text "Trombolisis_intr" with a yellow box labeled "rTPA" underneath it, followed by "Hora_primer_bolus_trombolisis_rtPA" with a green box labeled "a las 13.33 h" underneath it. The two grey boxes overlap, illustrating overlapping mentions.

Same tags, same normalized times, overlapping mentions (note the missing *h*), **OK**

Results

| Model | Examples | Tp | Fp | Fn | Acc | Pre | Rec | F1 |
|------------|----------|------|-----|-----|-------|-------|-------|--------------|
| Biomedical | 5455 | 5125 | 675 | 330 | 0.836 | 0.884 | 0.940 | 0.911 |
| Clinical | 5455 | 5104 | 638 | 351 | 0.838 | 0.889 | 0.936 | 0.912 |

Global average results comparing Biomedical and Clinical models



Differences between Biomedical and Clinical models.

Results

| tag | examples | Biomedical | Clinical | diff |
|---|-------------|--------------|--------------|---------------|
| NIHSS | 786 | 0.847 | 0.830 | 1.7 |
| TAC_craneal | 625 | 0.983 | 0.983 | 0.0 |
| mRankin | 275 | 0.961 | 0.966 | -0.5 |
| SECCION_EXPLORACIONES_COMPLEMENTARIAS | 217 | 0.954 | 0.954 | 0.0 |
| SECCION_MOTIVO_DE_INGRESO | 203 | 0.983 | 0.974 | 0.9 |
| SECCION_TRATAMIENTO_Y_RECOMENDACIONES_AL_ALTA | 200 | 0.980 | 0.990 | -1.0 |
| SECCION_PROCESO_ACTUAL | 185 | 0.981 | 0.981 | 0.0 |
| SECCION_EXPLORACION_FISICA | 177 | 0.969 | 0.972 | -0.3 |
| SECCION_TRATAMIENTO_HABITUAL | 155 | 0.937 | 0.950 | -1.3 |
| SECCION_EVOLUCION | 147 | 0.964 | 0.961 | 0.3 |
| Trombolisis_intravenosa | 146 | 0.925 | 0.906 | 1.9 |
| SECCION_ANTECEDENTES | 137 | 0.974 | 0.962 | 1.2 |
| SECCION_EXPLORACION_FISICA_DURANTE_HOSPITALIZACION | 136 | 0.842 | 0.803 | 3.9 |
| Trombectomia_mecanica | 122 | 0.799 | 0.816 | -1.7 |
| SECCION_EXPLORACION_FISICA_EN_URGENCIAS | 121 | 0.886 | 0.836 | 5.0 |
| Ictus_isquemico | 117 | 0.895 | 0.896 | -0.1 |
| SECCION_DESTINO_AL_ALTA | 113 | 0.968 | 0.968 | 0.0 |
| Etiologia | 110 | 0.854 | 0.861 | -0.7 |
| SECCION_DIAGNOSTICOS | 110 | 0.943 | 0.960 | -1.7 |
| ASPECTS | 107 | 0.811 | 0.869 | -5.8 |
| SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_PLANTA_DE_NEUROLOGIA | 104 | 0.755 | 0.813 | -5.8 |
| Tratamiento_antiagregante | 93 | 0.882 | 0.875 | 0.7 |
| SECCION_ANTECEDENTES_PATOLOGICOS | 91 | 0.937 | 0.937 | 0.0 |
| SECCION_TRATAMIENTO_AL_ALTA | 91 | 0.941 | 0.945 | -0.4 |
| Tratamiento_anticoagulante | 86 | 0.667 | 0.687 | -2.0 |
| SECCION_EXPLORACIONES_COMPLEMENTARIAS_EN_URGENCIAS | 75 | 0.938 | 0.938 | 0.0 |
| SECCION_SITUACION_FUNCIONAL | 72 | 0.966 | 0.973 | -0.7 |
| Arteria_afectada | 64 | 0.756 | 0.764 | -0.8 |
| Lateralizacion | 59 | 0.875 | 0.850 | 2.5 |
| SECCION_TIPO_DE_INGRESO | 53 | 0.981 | 0.991 | -1.0 |
| SECCION_PROCEDIMIENTOS | 50 | 0.961 | 0.980 | -1.9 |
| SECCION_EXPLORACION_FISICA_AL_ALTA | 46 | 0.893 | 0.893 | 0.0 |
| SECCION_ANTECEDENTES_PERSONALES | 42 | 0.988 | 0.988 | 0.0 |
| SECCION_RECOMENDACIONES | 42 | 0.864 | 0.930 | -6.6 |
| SECCION_MOTIVO_DEL_ALTA | 40 | 1.000 | 1.000 | 0.0 |
| SECCION_ANTECEDENTES_QUIRURGICOS | 32 | 0.853 | 0.870 | -1.7 |
| Localizacion | 31 | 0.703 | 0.714 | -1.1 |
| SECCION_CONTROL | 31 | 0.921 | 0.984 | -6.3 |
| Test_de_disfagia | 27 | 1.000 | 1.000 | 0.0 |
| Hora_TAC | 22 | 0.750 | 0.842 | -9.2 |
| Tiempo_puerta_aguja | 18 | 0.955 | 0.978 | -2.3 |
| Hora_primer_bolus_trombolisis_rtPA | 18 | 0.895 | 0.919 | -2.4 |
| Hemorragia_cerebral | 17 | 0.684 | 0.743 | -5.9 |
| Hora_recanalizacion | 10 | 0.667 | 0.909 | -24.2 |
| SECCION_ANTECEDENTES_FAMILIARES | 10 | 0.909 | 0.824 | 8.5 |
| SECCION_DIAGNOSTICO_PRINCIPAL | 9 | 0.842 | 1.000 | -15.8 |
| Hora_inicio_trombectomia | 9 | 1.000 | 0.941 | 5.9 |
| SECCION_DIAGNOSTICOS_SECUNDARIOS | 8 | 0.941 | 0.533 | 40.8 |
| Ataque_isquemico_transitorio | 6 | 0.615 | 0.769 | -15.4 |
| Hora_primera_serie_trombectomia | 5 | 0.600 | 0.600 | 0.0 |
| Hora_fin_trombectomia | 5 | 1.000 | 1.000 | 0.0 |
| ALL | 5455 | 0.911 | 0.912 | -0.097 |

Conclusions



- The information extraction task was complex and ambitious, with 51 different types of variables.
- In most cases, the variables are '**context-dependent**', which adds an extra difficulty of the task.
- **Temporal variables** are a case apart: in most cases the textual evidence shows an enormous variety. Such is the variety that, for the pre-annotation tool, we decided not to address the coding of these variables and limited ourselves to coding dates and times without going any further.
- The model managed to learn complex aspects such as the '**context sensitivity**' (this is very clear in the diagnostic variables, for example).
- The model managed to successfully learn the **complex temporal variables** that we had given up in the rule based system.

| Time variable | Examples | Biomedical | Clinical |
|------------------------------------|----------|------------|----------|
| Hora_TAC | 22 | 0.750 | 0.842 |
| Tiempo_puerta_aguja | 18 | 0.955 | 0.978 |
| Hora_primer_bolus_trombolisis_rtPA | 18 | 0.895 | 0.919 |
| Hora_recanalizacion | 10 | 0.667 | 0.909 |
| Hora_inicio_trombectomia | 9 | 1.000 | 0.941 |
| Hora_primera_serie_trombectomia | 5 | 0.600 | 0.600 |
| Hora_fin_trombectomia | 5 | 1.000 | 1.000 |

Conclusions

- We did not demonstrate that retraining with clinical data improves the model. We believe that
 - (i) more clinical data (from the stroke domain) could have improved the system,
 - (ii) mixing data from the very beginning would have positive effect, we are working on this
- **The results of the deep learning models are pretty good, reaching 91% F1 on average.** That is without applying any other (post)-process for system improvement. In this exercise we just wanted to evaluate the performance of deep learning techniques.
- **The results obtained demonstrate that the use of language technologies can be of great help in clinical information extraction tasks, as in the case of ICTUSnet.**

Demo & prototype

ICTUSnet

Cargar informes

Bandeja de entrada

Completados

Todos los informes

321108781.utf8

321687159.utf8

324602237.utf8

325139862.utf8

328077361.utf8

328342806.utf8

328359837.utf8

330011073.utf8

330459779.utf8

330682083.utf8

Procedimientos II - trombectomía

Tratamientos

Pruebas y escalas de valoración

TAC craneal

Fecha: dd/mm/aaaa

Hora: --:--

ASPECTS

ASPECTS

mRankin

Previsión: 0

Al alta: 2

NIHSS

Previsión: 0

Al alta: 11

Como pruebas complementarias, se realizó estudio neurosonológico de control que mostró a nivel de arteria carótida interna terminal derecha región con flujo aliasing y aceleración hasta 168cm/seg compatible con estenosis leve contra patrón TIBI IV. Se realizó ecocardiograma transtorácico que mostró aurícula izquierda, raíz aórtica y aorta ascendente ligeramente dilatadas y alteración diastólica de la relajación, sin otras alteraciones significativas. En los estudios neurosonológicos de control inicialmente se observó un patrón TIBI IV, y dos días después, simetría de ambas ACMs y ambos sifones carotídeos, además de un estudio de shunt de óptima calidad que resultó negativo. Una analítica con hemograma, hemostasia, perfil lipídico, magnesio, calcio, fosfato, hemoglobina glicada, TSH y proteinograma que no muestra alteraciones. Los marcadores tumorales (CA 15.3, CA 19.9, CEA, aFP y PSA) son bajos o normales. La monitorización electrocardiográfica continua durante los 4 días de ingreso no evidencia arritmias embolígenas.

Se termina orientando como ictus isquémico de territorio profundo de ACM derecha de etiología indeterminada tras estudio completo y mecanismo probablemente embólico (posibilidad de embolismo recanalizado versus aterotrombótico) clínicamente síndrome de alarma capsular motor derecho, tratado con rTPA. A pesar de la normalización del patrón DTC que va a favor de embolo recanalizado al estar pendiente de estudio RNM esta próxima semana se decide mantener la doble antiagregación y estatinas a dosis altas hasta ver resultado para modificarlo retirando uno de los antiagregantes y ajustando o estatinas en la visita de CCEE de control después de la RNM.

Al alta: Barthel = 100
NIHSS = 0
mRS = 2

<http://temu.bsc.es:81/>

[/ICTUSnet_time_variables_and_gs/normalized_times_test_predictions_brat/323767062.utf8](http://temu.bsc.es:81/)

SECCION TRATAMIENTO HABITUAL

42 MEDICACIÓN HABITUAL

43 CO-VALS FORTE 160MG/25MG VALSARTAN+DIURETIC 1 x 24 h.

44 Indefinida EMCONCOR COR 2,5MG BISOPROLOL, FUMARAT 1 x 24 h.

Tratamiento anticoagulante

45 Indefinida SINTROM 4MG ACENOCUMAROL 1 x 24 h.

46 Indefinida

SECCION PROCESO ACTUAL

49 PROCÉS ACTUAL

50 Paciente de 82 años que el día 11/01 es encontrada por familiares a las 19.40h en domicilio acostada en la cama con disminución nivel consciencia y debilidad hemicuerpo derecho por lo que se avisa al SEM quien objetiva RACE 8 y activa código ictus desde domicilio a las 20.15h.

51 La paciente estaba vestida con ropa de cama (ella siempre va a la cafetería a desayunar y hoy no la han visto, por lo que suponemos que es ictus del despertar).

52 La última vez vista asintomática el 10/01 al mediodía.

54 A su llegada al HUB a las 21.02h, TA 144/121mmHg, Coagulchek 1.5.

55 A nivel NRL se encuentra estuporosa (mueca y localiza al dolor) (2), orientación no valorable por mutismo (2+3), no cumple órdenes (2), miosis bilateral hiporreactiva, r.

56 oculocefálicos horizontales presentes, r, corneal derecho debil, HH D(2), PFSN D (2), hemiplejía derecha B 0/5con espasticidad (4), C 2/5 (3), sensibilidad no valorable (1), RCP extensor derecho, flexor izquierdo. NIHSS 21.

57 Frialdad 1-2-mitad 3er dedo mano derecha parética que recupera en unas horas.

TAC craneal

60 Se realiza TC craneal en el que se objetiva infarto de todo el territorio de la ACM izquierda con transformación hemorrágica asociada a nivel de GGBB izquierdos abierto a ventriculos, e infarto de pequeño territorio de ACA izquierda.

61 Sin desplazamiento línea media.

SECCION TRATAMIENTO HABITUAL

42 MEDICACIÓN HABITUAL

43 CO-VALS FORTE 160MG/25MG VALSARTAN+DIURETIC 1 x 24 h.

44 Indefinida EMCONCOR COR 2,5MG BISOPROLOL, FUMARAT 1 x 24 h.

Tratamiento anticoagulante

45 Indefinida SINTROM 4MG ACENOCUMAROL 1 x 24 h.

46 Indefinida

SECCION PROCESO ACTUAL

49 PROCÉS ACTUAL

50 Paciente de 82 años que el día 11/01 es encontrada por familiares a las 19.40h en domicilio acostada en la cama con disminución nivel consciencia y debilidad hemicuerpo derecho por lo que se avisa al SEM quien objetiva RACE 8 y activa código ictus desde domicilio a las 20.15h.

51 La paciente estaba vestida con ropa de cama (ella siempre va a la cafetería a desayunar y hoy no la han visto, por lo que suponemos que es ictus del despertar).

52 La última vez vista asintomática el 10/01 al mediodía.

54 A su llegada al HUB a las 21.02h, TA 144/121mmHg, Coagulchek 1.5.

55 A nivel NRL se encuentra estuporosa (mueca y localiza al dolor) (2), orientación no valorable por mutismo (2+3), no cumple órdenes (2), miosis bilateral hiporreactiva, r.

56 oculocefálicos horizontales presentes, r, corneal derecho debil, HH D(2), PFSN D (2), hemiplejía derecha B 0/5con espasticidad (4), C 2/5 (3), sensibilidad no valorable (1), RCP extensor derecho, flexor izquierdo. NIHSS 21.

57 Frialdad 1-2-mitad 3er dedo mano derecha parética que recupera en unas horas.

TAC craneal

60 Se realiza TC craneal en el que se objetiva infarto de todo el territorio de la ACM izquierda con transformación hemorrágica asociada a nivel de GGBB izquierdos abierto a ventriculos, e infarto de pequeño territorio de ACA izquierda.

https://temu.bsc.es/ICTUSnet/diff.xhtml?diff=%2FICTUSnet_time_variables_and_gs%2Ftest_brat_gs%2F#/ICTUSnet_time_variables_and_gs/normalized_times_test_prediction_s_brat/323767062.utf8

#ICTUSnetBCN2020



THANK YOU

