# Using natural language processing techniques to study and regulate emergency department flows : development and application to the study of trauma risks based on ED venues in Bordeaux

Gabrielle Chenais

▶ **To cite this version:**

Gabrielle Chenais. Using natural language processing techniques to study and regulate emergency department flows : development and application to the study of trauma risks based on ED venues in Bordeaux. Human health and pathology. Université de Bordeaux, 2023. English. NNT : 2023BORD0205 . tel-04554258

**HAL Id: tel-04554258**
**https://theses.hal.science/tel-04554258**

Submitted on 22 Apr 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

université de BORDEAUX

THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE

DOCTEUR.E DE
L'UNIVERSITÉ DE BORDEAUX

**ECOLE DOCTORALE SOCIÉTÉS, POLITIQUE, SANTÉ PUBLIQUE**
**SANTÉ PUBLIQUE INFORMATIQUE ET SANTÉ**

Par **Gabrielle CHENAIS**

# EXPLOITATION DES TECHNIQUES DE TRAITEMENT AUTOMATIQUE DU LANGAGE POUR L'ETUDE ET LA REGULATION DES FLUX AUX URGENCES

## Développement et application à l'étude des risques de traumatismes à partir des admissions aux urgences à Bordeaux.

*USING NATURAL LANGUAGE PROCESSING TECHNIQUES TO STUDY AND REGULATE EMERGENCY DEPARTMENT FLOWS*
*Development and application to the study of trauma risks based on ED venues in Bordeaux.*

Sous la direction de : **Emmanuel LAGARDE**

Soutenue le 11 septembre 2023

Membres du jury :

| | | | | | |
|---|---|---|---|---|---|
| Mme | Névéol, | Aurélie | Maître de conférence, HDR | CNRS | Rapporteur |
| M. | Darmoni | Stefan | PU-PH | Université de Rouen | Rapporteur |
| M. | Jouhet | Vianney | PU-PH | Université de Bordeaux | Examinateur |
| M. | Gourraud | Pierre-Antoine | PU-PH | Université de Nantes | Président |
| Mme | Beltzer | Nathalie | Dr | Santé Publique France | Examinateur |
| M. | Lagarde | Emmanuel | HDR | Université de Bordeaux | Invité |

# ABSTRACT

The TARPON (Traitement Automatique des Résumés de Passage aux urgences dans le but de créer un Observatoire National du traumatisme) project aims to demonstrate the feasibility of setting up a French observatory of trauma. Emergency Departments (EDs) generate a large volume of health-related data and approximately one-third of ED visits are the result of trauma. Most of the information contained in electronic health records is in the form of free text format and manual information extraction is time and resource consuming. Artificial Intelligence (AI) and particularly Natural Language Processing (NLP) could optimize this process. NLP has seen a recent breakthrough with the introduction of deep learning and in particular the Transformer architecture. These Large Language Models (LLMs) have reached the state-of-the-art for most NLP tasks and their use for clinical and medical data is promising. To explore the potential of Transformers for trauma classification (multi-class), we conducted an evaluation using free-text clinical notes from a single large University Hospital (Bordeaux) ED. A total of 69,110 free-text clinical notes generated between 2012 and 2019 were manually annotated, with 22,481 identified as traumas. To compare the performance of traditional machine learning classifiers and Transformer models, we employed different architectures (BERT and GPT-2), varied sizes, pre-training corpora languages and tokenizers (OSCAR, Wiki, and CCNET). Additionally, we investigated the impact of incorporating a pre-training step on a domain-specific corpus. Our findings revealed that bagging algorithms and Light Gradient Boosting exhibited similar results to the lower-performing Transformers. Interestingly, we discovered that larger models did not necessarily translate to better performance, but the choice of pre-training corpora significantly influenced the outcomes. The best results, with an average F1-score of 0.976, were achieved using a GPT-2 architecture with two steps of pre-training utilizing a French corpus then with a domain-specific corpus. These results highlight the potential of Transformers, particularly when an unsupervised pre-training with a domain-specific corpus is performed, in the accurate classification of traumas based on free-text clinical notes.

Our contribution to the TARPON project laid the groundwork for the use of LLMs for processing clinical notes. These models, which are becoming increasingly efficient and powerful, have led to a recent paradigm shift in NLP. Most AI applications currently in use in emergency medicine are based on NLP and automatic speech recognition because of the privileged documentation medium of free or semi-structured text or the practitioner-patient interaction. However, these applications lack proper derivation, validation, or impact evaluations that are performed rigorously and independently. Building a trustworthy, safe, and explainable AI requires a holistic approach that encompasses all sociotechnical aspects involved. Human factors such as participatory design and multi-stakeholder approaches are important for building such AI systems. Inclusiveness begins at the very beginning of the design step, with the inclusion of stakeholders. All possible biases and risks should be identified and documented before any initiation, and they should be monitored continuously. However, when emergency medicine is concerned with the development of AI applications, several principles mentioned above collide, and trade-offs must be determined. How can we determine the trade-off among interpretability and performance, time, and explainability?

How can transparency be ensured when intellectual property is involved? How can liability be determined when AI harms?
To ensure the safety of patients, healthcare professionals and researchers, we need to bring together all the stakeholders involved in the development of such healthcare tools. Legislators, decision-makers, insurers and public authorities have a duty to work together to provide the best possible support for a change that is taking place in spite of them.

**Keywords** : Artificial Intelligence, Natural Language Processing, Transformer, Emergency, Trauma, Public Health Surveillance

# RÉSUMÉ

Le projet TARPON (Traitement Automatique des Résumés de Passage aux urgences dans le but de créer un Observatoire National du traumatisme) vise à démontrer la faisabilité de la mise en place d'un observatoire français des traumatismes. Les services d'urgences génèrent un volume important de données de santé et environ un tiers des visites aux urgences sont liées à des traumatismes. La plupart des informations contenues dans les dossiers médicaux électroniques sont sous forme de texte libre, et l'extraction manuelle d'informations est une tâche chronophage qui nécessite beaucoup de ressources. L'intelligence artificielle (IA) et plus particulièrement le traitement automatique du langage naturel (TALN) pourraient optimiser ce processus. Le TALN a connu un changement de paradigme récent avec l'introduction de l'apprentissage profond et en particulier l'architecture de type Transformer. Ces larges modèles de langage (LLMs) ont atteint l'état de l'art pour la plupart des tâches de TALN et leur utilisation pour les données cliniques et médicales est prometteuse.

Afin d'explorer le potentiel des Transformers dans la classification multi-classe des traumatismes, nous les avons évalués sur des notes cliniques en texte libre provenant d'un centre hospitalier universitaire (Bordeaux). Un total de 69 110 notes cliniques en texte libre générées entre 2012 et 2019 ont été annotées manuellement, parmi elles, 22 481 ont été identifiées comme des traumatismes. Nous avons comparé les performances d'outils de classification issus du machine learning traditionel à des modèles de type Transformer. Concernant ces derniers, nous avons utilisé différentes architectures (BERT et GPT-2), des tailles de modèles variables et des modèles pré-entrainés avec des langues et des tokenizers différents pour les corpus de pré-entraînement (OSCAR, Wiki et CCNET). De plus, nous avons étudié l'impact de l'ajout d'une étape de pré-entraînement sur la base de données non labelisée des urgences. Les algorithmes de bagging et le Light Gradient Boosting ont obtenu des résultats similaires aux Transformers les moins performants. De plus, nous avons découvert que des modèles plus grands n'induisaient pas nécessairement par de meilleures performances, en revanche, le choix des corpus de pré-entraînement influençait les performances en classification. Les meilleurs résultats, avec un score F1 moyen de 0,976, ont été obtenus avec une architecture de type GPT-2 comprenant deux étapes de pré-entraînement non supervisé utilisant un corpus français puis la base de données entière. Ces résultats mettent en évidence la capacité des Transformers, en particulier lorsqu'un pré-entraînement non supervisé avec un corpus spécifique au domaine est effectué, dans la classification précise des traumatismes à partir de notes cliniques en texte libre.

Notre contribution au projet TARPON a posé les bases de l'utilisation des LLM pour la classification des notes cliniques. Ces modèles, de plus en plus efficaces et puissants, ont récemment entraîné un changement de paradigme dans le domaine du TALN. La plupart des applications d'IA actuellement utilisées en médecine d'urgence sont basées sur le TALN et la reconnaissance vocale automatique en raison mode de documentation privilégié (texte libre ou semi-structuré) des professionnels de santé ou de l'interaction entre le praticien et le patient. Cependant, ces applications ne bénéficient pas d'études de validation et de dérivation ou d'évaluations d'impact adéquates et effectuées de manière rigoureuse et indépendante. La construction d'une IA fiable, sûre et explicable nécessite une approche holistique englobant tous les aspects sociotechniques impliqués. Des facteurs humains tels que la conception

participative et les approches transversales sont importants pour la construction de tels systèmes d'IA Dès le commencement de l'étape de conception, il est essentiel d'adopter une approche inclusive en impliquant activement toutes les parties prenantes. Il est impératif d'identifier et de documenter tous les biais et risques potentiels avant de déployer une IA, et de les surveiller de manière continue par la suite.
Néanmoins, lorsqu'il s'agit du développement d'applications d'IA en médecine d'urgence, divers principes mentionnés précédemment entrent en conflit, nécessitant ainsi l'établissement de compromis. Comment pouvons-nous trouver un juste équilibre entre interprétabilité et performance, le facteur temps lié à l'urgence (parfois vitale) et l'explicabilité ? Comment assurer la transparence lorsqu'il y a des enjeux de propriété intellectuelle ? Comment déterminer la responsabilité en cas de préjudice causé par l'IA ?
Aussi afin de garantir la sécurité des patients et des professionnels de santé, mais aussi des chercheurs, il convient de fédérer tous les acteurs impliqués dans le développement de tels outils en santé. Les législateurs, les instances décisionnaires, les assureurs et les pouvoirs publics ont le devoir de s'unir pour accompagner au mieux un changement qui est en train de se passer, malgré eux.

**Mots clés** : Intelligence Artificielle, Traitement Automatique du Langage, Transformer, GPT, Urgences, Traumas, Surveillance, Santé Publique

# REMERCIEMENTS

A mon directeur de thèse, Pr. Emmanuel Lagarde, Merci pour ton accompagnement, tes idées novatrices qui repoussent les limites du NLP pour les données de santé et ta détermination concernant ce projet qui a encore des promesses à tenir.

Au Professeur Pierre-Antoine Gourraud, Merci de présider le jury de cette thèse. Soyez assuré de ma reconnaissance pour votre apport lors de cette soutenance.

Aux rapporteurs, Pr. Aurélie Névéol et Pr. Stefan Darmoni, C'est un réel honneur d'avoir soumis mon travail à votre expertise. Je n'aurais pas pu espérer mieux et je suis heureuse que vous ayez accepté de porter ce travail.

Aux autres membres du jury, le Dr. Vianney Jouhet et le Dr. Beltzer. Je vous remercie d'avoir accepté de participer à mon jury de thèse, d'avoir pris le temps de lire ce document et de vous être intéressés à mes travaux de recherche.

Au Professeur Rodolphe Thiebaut, je te remercie de m'avoir fait confiance et de m'avoir choisie parmi les candidats français pour le Master en public health data science. Tu as su voir mon potentiel et ma détermination sans failles et sans cela ce doctorat n'aurait pas été possible.

Aux membres de mon comité de suivi de thèse, Pr. Emmanuel Bacry, Pr. Gayo Diallo et Dr. Anne Gallay, Je tiens à vous exprimer ma gratitude dans cet accompagnement au cours de ces années de doctorat marquées par une pandémie et certains changements. Votre apport structuré et bienveillant aura permis que cette thèse se déroule dans les meilleures conditions possibles.

A l'équipe AHead, je ne saurais jamais assez vous exprimer toute ma gratitude, en particulier Antoine, Gayo, Cédric, Éric, Dylan, Hélène, Benjamin et Alexandre. J'ai découvert avec vous une réelle ouverture d'esprit, une stimulation intellectuelle sans faille, une ambiance professionnelle bienveillante et valorisante où les qualités de chaque membre sont mises en lumière pour créer une véritable émulation. Vous m'avez convaincue qu'il existe des milieux professionnels sains.

A Marie-Odile, ta bienveillance et ton empathie auront concouru à ce que ce doctorat se passe au mieux. Tu auras été un phare et je te suis très reconnaissante.

A tou.te.s les infirmière.e.s des urgences qui ont participé de près ou de loin au projet TARPON. Votre investissement et votre dynamisme aura permis de constituer une base de données labelisée de qualité.

Aux membres du Heath Data Hub, Laureen Majed, Jade Viarigi, Metty Mavounia, Emmanuel Bacry et Stéphanie Combes, Je vous remercie pour cette collaboration riche, innovante et instructive. Harold, notre serveur est né grâce à vous.

Aux stagiaires de Master 2 que j'ai pu accompagner, Melissa et Chloé, je vous remercie de m'avoir aidée à grandir et de m'avoir permise de continuer à exercer cette pédagogie qui m'est si chère.

Aux frenchies de sci-kit learn et Hugging face.

A ma famille, mes enfants, amis, et à toi.

# TABLE OF CONTENT

LIST OF ACRONYMS ..... 1
LIST OF FIGURES ..... 3
LIST OF TABLES ..... 6
I. GENERAL CONTEXT ..... 7
I.1 Traumas and Injury: a Global and Public Health Burden ..... 7
I.1.1 Definitions ..... 7
I.2 Injury: A Leading Cause of the Global Burden of Disease ..... 8
I.2.1 French Epidemiology of Traumas ..... 8
I.2.2 Trauma Prevention ..... 9
I.3 Building a National Trauma Surveillance System ..... 10
I.3.1 Existing trauma related surveillance systems in France ..... 10
I.3.2 French Injury Surveillance System: a political will and an ongoing project ..... 12
I.3.3 Injury surveillance system requirements ..... 13
I.3.4 French emergency surveillance system ..... 15
I.3.5 From Electronic Health Record to ED visit summaries for trauma ..... 17
II. NATURAL LANGUAGE PROCESSING FOR CLINICAL DATA ..... 19
II.1 Narrative Clinical data ..... 19
II.1.1 Natural Language and Sub-languages ..... 20
II.1.2 Natural Language Processing for French Clinical Textual Data ..... 21
II.2 Natural Language Processing for Text Classification ..... 24
II.2.1 Setting the Frame ..... 24
II.2.2 Document representation and feature selection ..... 27
II.2.3 Statistical and Traditional Machine Learning classification algorithms ..... 39
II.2.4 Deep Learning versus Statistical and Traditional Machine Learning ..... 45
II.2.5 Artificial Neural Networks ..... 46
II.2.6 Data Augmentation in NLP ..... 75
II.2.7 Language Model Evaluation ..... 78
III. NATURAL LANGUAGE PROCESSING FOR PUBLIC HEALTH SURVEILLANCE: THE TARPON PROJECT ..... 82
III.1 TARPON: Context ..... 82
III.1.1 Project aim ..... 82
III.2 TARPON: Methods ..... 82
III.2.1 Medical ethics regulations and GDPR ..... 82

III.2.2 Database ..... 82
III.2.3 Exploratory text analysis ..... 83
III.2.4 Labeling strategy ..... 85
III.2.5 Models and experiment settings ..... 87
III.2.6 Self-supervised learning and Fine-tuning phase ..... 93
III.2.7 Test phase ..... 93
III.2.8 Labeled datasets ..... 93
III.2.9 Error analysis ..... 95
III.3 TARPON: Results ..... 95
III.3.1 Clinical notes' structure ..... 95
III.3.2 Linguistic features ..... 96
III.3.3 Topic Modeling ..... 98
III.3.4 Fine-tuning performance of models ..... 100
III.3.5 Performance of models ..... 101
III.3.6 Error analysis ..... 104
III.4 TARPON: Discussion ..... 107
III.4.1 Transformers: a new state of the art ..... 107
III.4.2 Self-supervised training on domain specific corpus and tokenizer ..... 108
III.4.3 Taxonomy ..... 108
III.5 TARPON: Conclusion ..... 109
III.6 TARPON: Perspectives ..... 109
III.6.1 Improvement of the annotation grid ..... 109
III.6.2 Future epidemiological steps of the TARPON project ..... 110
III.6.3 Towards a French trauma surveillance system ..... 110
IV. ARTIFICIAL INTELLIGENCE IN EMERGENCY MEDICINE: VIEWPOINT OF CURRENT APPLICATIONS AND FORESEEABLE OPPORTUNITIES AND CHALLENGES ..... 112
IV.1 Flow Challenges in Emergency Departments ..... 112
IV.1.1 Factors of Emergency Departments crowding ..... 112
IV.1.2 Consequences of Emergency Departments crowding ..... 113
IV.2 ARTIFICIAL INTELLIGENCE : A POSSIBLE SOLUTION ..... 113
IV.2.1 Artificial Intelligence in Emergency Medicine: Current Applications and Foreseeable Opportunities ..... 113
IV.2.2 ED and EMD data processing enhanced by AI for public health surveillance ..... 122

IV.3 CHALLENGES POSED BY ARTIFICIAL INTELLIGENCE FOR EMERGENCY MEDICINE AND PUBLIC HEALTH SURVEILLANCE ..... 123
IV.3.1 Ethical and legal challenges posed by the implementation of Artificial Intelligence in Emergency Medicine ..... 123
IV.3.2 Safety, fairness and bias management ..... 124
IV.3.3 Transparency, Accountability and Liability ..... 129
IV.3.4 Explainability and Interpretability ..... 130
IV.3.5 Autonomy ..... 131
IV.3.6 Privacy-Enhanced ..... 132
IV.4 Technical challenges ..... 132
V.2.1 Training and data challenges ..... 132
IV.4.1 Integration Into Routine Clinical Workflow ..... 133
IV.5 Conclusion ..... 134
V. GENERAL CONCLUSION ..... 135
VI. BIBLIOGRAPHY ..... 136
VII. PUBLICATIONS ..... 161
VIII. APPENDIX ..... 162

# LIST OF ACRONYMS

| | |
|---|---|
| AI | Artificial Intelligence |
| AIC | Akaike Information Criterion |
| ANN | Artificial Neural Networks |
| ARS | Agence Régionale de Santé |
| ASR | Automatic Speech Recognition |
| ATIH | Agence Technique de l'Information sur l'Hospitalisation |
| AUC | Area Under Curve |
| BERT | Bidirectional Encoder Representations from Transformers |
| BIC | Bayesian Information Criterion |
| BPE | Byte-Pair Encoding |
| BBPE | Byte-level Byte-Pair Encoding |
| BOW | Bag of Words |
| CART | Classification And Regression Trees |
| CCNet | Criss-Cross attention for semantic segmentation |
| CDSS | Criss-Cross attention for semantic segmentation |
| CHAID | Chi-squared Automatic Interaction Detector |
| CNN | Convolutional Neural Network |
| DL | Deep Learning |
| E2E | End-to-End |
| ED | Emergency Department |
| EM | Emergency Medical Dispatch |
| EMT | Emergency Medical Technician |
| EHR | Electronic Health Record |
| ER | Emergency Room |
| FlauBERT | French Language Understanding via Bidirectional Encoder Representations from Transformers |
| FFNN | Feed-Forward Neural Networks |
| GloVe | Global Vectors for word representation |
| GP | General Practitioner |
| GPT | Generative Pre-trained Transformer |
| GPU | Graphical Per Unit |
| HDH | Health Data Hub |
| ICD-10 | International Classification of Diseases v10 |
| ICECI | International Classification of External Causes of Injury |
| ICU | Intensive Care Unit |
| IE | Information Extraction |
| INVS | Institut National de Veille Sanitaire |
| KNN | K-Nearest Neighbor |
| LDA | Latent Dirichlet Allocation |
| LLM | Large Language Model |
| LSA | Latent Semantic Analysis |
| LSTM | Long Short-Term Memory |
| MAD | Median Absolute Deviation |
| MCC | Matthews Correlation Coefficient |
| MDS | Minimum Data SET |
| ML | Machine Learning |

| | |
|---|---|
| MLM | Masked Language Model |
| MLP | Multi-Layer Perceptron |
| MAUP | Modifiable Areal Unit Problem |
| MVA | Motor Vehicle Accident |
| NLP | Natural Language Processing |
| NSO | National Suicide Observatory |
| ODS | Optional Data Set |
| OHCA | Out of Hospital Cardiac Arrest |
| OSCAR | Open Super-large Crawled Aggregated coRpus |
| OSCOUR | Organisation de la Surveillance Coordonnée des Urgences |
| OOV | Out Of Vocabulary |
| ORU | Observatoire Régional des Urgences |
| PFM | Pre-trained Foundation Model |
| POS | Part-Of-Speech |
| PLSA | Probabilistic Latent Semantic Analysis |
| RTA | Road Traffic Accident |
| RNN | Recurrent Neural Network |
| RoBERTa | Robustly Optimized BERT Pretraining |
| TARPON | Traitement Automatique des Résumés de Passage aux Urgences dans le but de créer un Observatoire National des traumatismes |
| TF-IDF | Term-Frequency - Inverse Document Frequency |
| SD | Standard Deviation |
| SEDV | Summary of Emergency Department Visit |
| SMART | System for the Mechanical Analysis and Retrieval of Text |
| SPF | Santé Publique France |
| SVC | Support Vector Classification |
| SVM | Support Vector Machine |
| RCT | Randomized Controlled Trial |
| RNN | Recurrent Neural Network |
| RPU | Résumé de Passage aux Urgences |
| UHCD | Unité d'Hospitalisation de Courte Durée |
| UMAP | Uniform Manifold Approximation and Projection for Dimension Reduction |
| UTF-8 | Universal Character Set Transformation Format-8 bits |
| VSM | Vector Space Model |
| YLL | Years of Lost Life |

# LIST OF FIGURES

Figure I.1 Proportions of total deaths in France, 2017, regarding the injury causes[7] ..... 8
Figure I.2 Public health approach to injury prevention. ..... 9
Figure I.3 Building blocks of an Injury Surveillance System ..... 15
Figure I.4 Stakeholders involved in French Emergency Surveillance. ..... 16
Figure I.5 Example of a trauma-related clinical note extracted from Bordeaux University Hospital database. TC: Trauma Crânien, PCI: Perte de Connaissance Initiale, G: Gauche ..... 18
Figure II.1 Tasks and their prevalence in Wu et al. study from the NLP and the clinical perspective. ..... 21
Figure II.2 Growth of broad architectures in deep learning over the years. Wu et al.[56] ..... 23
Figure II.3 Data Mining Techniques ..... 25
Figure II.4 Natural Language Processing Pipeline ..... 27
Figure II.5 Euclidean distance illustration ..... 29
Figure II.6 Cosine similarity illustration ..... 30
Figure II.7 Two different ways of splitting a text with word tokenization ..... 36
Figure II.9 Character-level tokenization ..... 37
Figure II.10 Byte Pair Encoding with bigrams ..... 38
Figure II.11 Artificial Intelligence, Machine Learning and Deep Learning connections ..... 46
Figure II.12 Types of Artificial Neural Networks ..... 47
Figure II.13 Convolutional Neural Network[132] ..... 48
Figure II.14 Unrolled RNN[135] ..... 50
Figure II.15 Types of RNN ..... 51
Figure II.16 LSTM units with the 4 interacting layers[135] ..... 53
Figure II.17 Sigmoid layer of an LSTM cell[135] ..... 53
Figure II.18 Cell state of an LSTM[135] ..... 54
Figure II.19 Forget gate layer of LSTM[135] ..... 54
Figure II.20 Input gate of LSTM[135] ..... 55
Figure II.21 Output gate of LSTM[135] ..... 55
Figure II.22 Output gate of LSTM[135] ..... 55
Figure II.23 GRU structure $zt$ and $rt$ represent the update gate and reset gate respectively[140] ..... 56
Figure II.24 The transformer architecture proposed by Vaswani et al.[66] ..... 58
Figure II.25 Encoder/decoder components of the Transformer[66]. ..... 58
Figure II.26 Encoder component of the Transformer[142] ..... 59
Figure II.27 Decoder component of the Transformer[142] ..... 59
Figure II.28 Mapping of the words, their matching index ID and Embeddings[142] ..... 59
Figure II.29 Flow of the words in the bottom encoder[142] ..... 60
Figure II.30 Weights and Query/key/value matrix for each word[142] ..... 60
Figure II.31 Illustration of the dot product of query and key vectors[142] ..... 61
Figure II.32 Illustration of the production of the output of the self-attention layer[142] ..... 61
Figure II.33 Illustration of the matrix calculation of self-attention[142] ..... 62
Figure II.34 Multi-head attention of the Transformer[66] ..... 62
Figure II.35 Illustration of the separated weight matrices in Transformer[142] ..... 63
Figure II.36 Concatenation and linear normalization layers of the Transformer[66] ..... 63
Figure II.37. Illustration of the summary of the attention process[142] ..... 63

Figure II.38 Encoder-decoder attention layer of the Transformer[66] ..... 64
Figure II.39 Bottom decoder layer of the Transformer[66] ..... 64
Figure II.40 Illustration of the final linear and softmax layer[142] ..... 65
Figure II.41 BERTbase and BERTlarge size and architecture illustration (from Hugging face blog[179]) ..... 69
Figure II.42 Pre-training procedure of BERT. (from [67]) ..... 70
Figure II.43 BERT input representation (from [67]) ..... 70
Figure II.44 Decoder blocks of a decoder-only Transformer. The first decoder block Is expanded. (From[142]) ..... 72
Figure II.45 GPT architecture and training objectives (From[188]) ..... 73
Figure II.46 GPT-2 sizes and dimensionality (from[142]) ..... 74
Figure II.47 Methods of paraphrasing in text data augmentation ..... 75
Figure III.1 Composite variable type of trauma based on the annotation grid variables. ..... 86
Figure III.2 Example of a clinical note ..... 87
Figure III.3 Example of clinical notes ..... 93
Figure III.4 Distribution of the Number of Tokens per clinical notes categories. ..... 96
Figure III.5 Distribution of the major Part-Of-Speech tags (over 2%) normalized on length among clinical notes for both physicians and nurses' categories (french-camembert-postag-model's confidence scores are given upon each bar) ..... 98
Figure III.6 : Distribution of all the Part-Of-Speech tags normalized on length among clinical notes for both physicians and nurses' categories. ..... 98
Figure III.7 Top 5 topics identified by BERTopic with their most frequent words and scores ..... 99
Figure III.8 Hierarchical clustering of the top 50 topics identified by BERTopic ..... 99
Figure III.9 Example of clinical notes generated by GPTanam after self -supervised pre-training step. CT: Cranial Trauma, LOC: Loss Of Consciousness, MVA: Motor Vehicle Accident, LV: Light Vehicle ..... 100
Figure III.10 F1-score curves for CamemBERT-CCNET, FlauBERT-small, BelGPT2 and GPTanam on the validation dataset. ..... 101
Figure III.11 Confusion matrix of GPTanam model on the full test dataset ..... 104
Figure III.12 Plot of micro F1-scores of all models for each class for both the complete test dataset (blue bars) and the test dataset without potentially ambiguous content as regard to its classification (grey bars). ..... 106
Figure III.13 Confusion matrix of GPTanam model on the test dataset without ambiguous content ..... 107
Figure III.14 From "Harnessing the Power of LLMs in Practice: A Survey of ChatGPT and Beyond"[218] The evolutionary tree of modern LLMs traces the development of language models in recent years and highlights some of the most well-known models. ..... 111
Figure IV.1 AI business Landscape in Emergency Medicine in 2022 ..... 114
Figure IV.2 The emergency patient journey and where Artificial Intelligence is making or can make an impact. ..... 115
Figure IV.3 Gradient boosting explanation ..... 120
Figure IV.4 The human digital twin ..... 121
Figure IV.5 Lifecycle and Key Dimensions of an AI System. National Institute of Standards and Technology (NIST) [299] ..... 123
Figure IV.6 AI actors across AI lifecycle stages. Note that AI actors in the AI Model dimension are separated as a best practice, with those building and using the models separated from

those verifying and validating the models. TEVV: Test, Evaluation, Verification and Validation ..... 124
Figure IV.7 Misaligned goals in artificial intelligence (AI) ..... 124
Figure IV.8 A centralized-server approach to federated learning (From [397]) ..... 133

# LIST OF TABLES

Table II.1 Summary of Large Language Models (inspired from[178]). Models are chronologically presented in the table. ..... 68
Table II.2 BERTbase and BERTlarge size and architecture ..... 69
Table III.1 Labels Distribution among Train, validation and test dataset. MVA: Motor Vehicle Accident ..... 94
Table III.2 Train, validation and test dataset characteristics ..... 94
Table III.3 Availability of clinical notes in the TARPON database ..... 95
Table III.4 Average Document Length for both the complete set of notes and for notes excluding those with an outlier number of tokens (in parentheses) ..... 95
Table III.5 Vocabulary differences by category ..... 96
Table III.6 Parts-of-speech matching for each tag of the French Treebank dataset[218] ..... 97
Table III.7 Micro F1-scores for all classes and average F1-score for all models AE: Accident of Exposure (to Bodily Fluids), MVA: Motor Vehicle Accident ..... 102
Table III.8 Micro F1-scores for all classes and selected models with micro average F1-scores and macro average precision on the complete test dataset. ..... 103
Table III.9 Micro F1-*scores for all classes and selected models with micro average F1-scores and macro average precision on the test dataset without ambiguous content.* ..... 106
Table IV.1 Examples of potential legal outcomes related to artificial intelligence (AI) use in clinical practice[383] ..... 130

# I. GENERAL CONTEXT

## I.1 Traumas and Injury: a Global and Public Health Burden

With an estimated proportion of 30%, traumas represent of large portion of French Emergency Departments (ED) activity[1].

### I.1.1 Definitions

An injury, also known as physical trauma, is defined as the physical damage that results when a human body is suddenly or briefly subjected to intolerable levels of energy. It can be a bodily lesion resulting from an acute exposure to an energy exceeding the threshold of physiological tolerance in amount, or it can be an impairment of function resulting from a lack of one or several vital elements (i.e., air, water, warmth), as in drowning, strangulation or freezing. The time between exposure to the energy and the appearance of an injury is short. The energy causing an injury may be[2]:

- mechanical (e.g., an impact with a moving or stationary object, such as a surface, knife or vehicle)
- radiant (e.g., a blinding light or a shock wave from an explosion)
- thermal (e.g., air or water that is too hot or too cold)
- electrical
- chemical (e.g., a poison or an intoxicating or mind-altering substance such as alcohol or a drug)

In other words, injuries are the acute, physical conditions listed in Chapter XIX (Injury, poisoning, and certain other consequences of external causes) and Chapter XX (External causes of morbidity and mortality) in the International Statistical Classification of Diseases and Related Health Problems, Tenth revision(ICD-10)[3].

The most common events causing injuries are[4]:

- interpersonal violence and sexual abuse;
- collective violence including wars, civil insurrections and riots;
- traffic collisions and
- incidents at home, at work and while participating in sports and other recreational activities

Injuries can occur in every environment from homes to the workplace, recreational settings including sports settings, and in transportation settings between these multiple environments. They can be classified in different ways such as by cause (intentional, accidents), by modality, by location and/or by activity[5].

## I.2 Injury: A Leading Cause of the Global Burden of Disease

Trauma represents a leading cause of mortality and morbidity worldwide. According to the Global Burden of Diseases study in 2017, injuries account for 8.02% (CI:7.74-8.17%) of deaths and 11.86% (CI:11.46-12.13%) of Years of Lost Lifes (YLLs) worldwide[6].

### I.2.1 French Epidemiology of Traumas

In 2017 in France, traumas and injuries accounted for 7.01% (CI:6.75- 7.33%) of deaths with a decreasing trend since 1990 while they represented 11.1% (CI:10.71- 11.47%) of the total of YLLs[7]. 34.87% (CI:33.37-36.3%) of total deaths in 2017 were attributable to injuries among 15 to 49-year-old French people.
Bege and al. showed in 2019 that patients admitted for trauma in French hospitals had a 5.9% mortality rate within 30-days. The patients' age and severity of injuries were strong predictors for mortality, while being female was a protective factor. They suggested that aging is a deleterious process in terms of mortality risk as a major increase of death rate was found among patients older than 75 years old[8]. The leading causes of death when people are injured are falls, unintentional injuries and road traffic accidents as can be seen on Figure I.*1*.

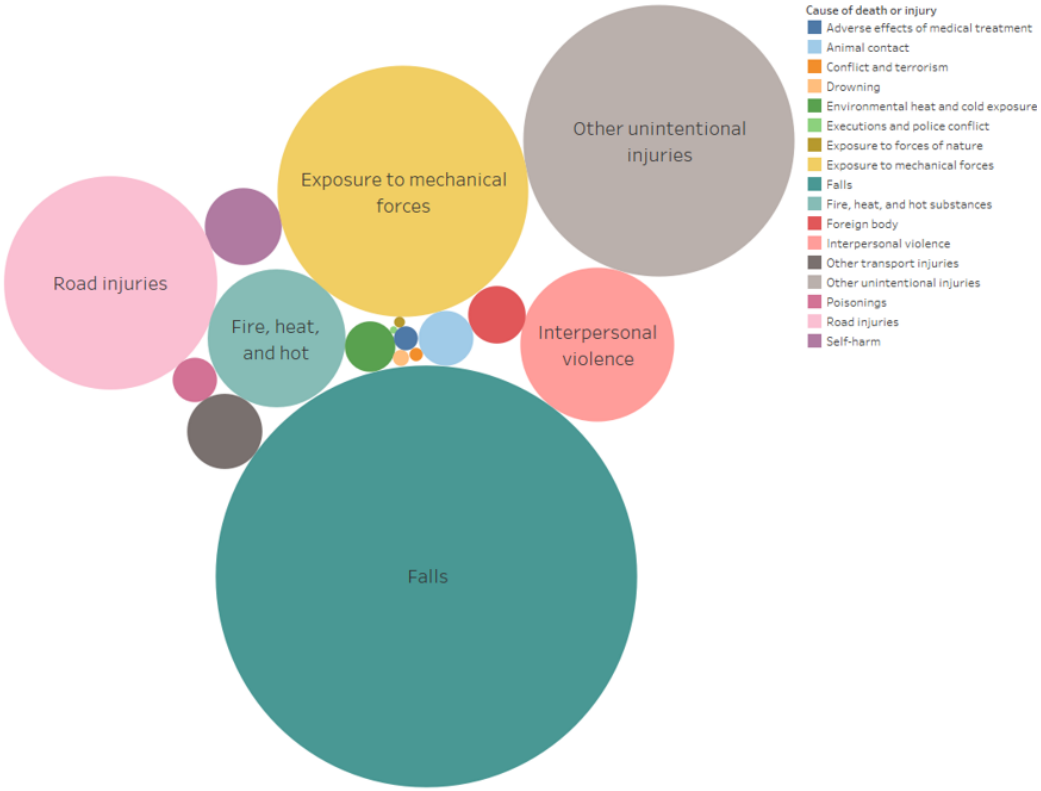



Figure I.1 Proportions of total deaths in France, 2017, regarding the injury causes[7]
The bubble sizes match the value of deaths percentages.

### I.2.2 Trauma Prevention

The scale of the injury problem is not a matter of dispute. As previously indicated, the number of people who die from injury every year runs into the millions. However, deaths are only a small part of the total injury problem; for each person killed, many more are seriously and permanently disabled, and even more are suffering from minor, short-term disabilities. The costs of injury mortality and morbidity are immense, not only in terms of lost economic opportunity and demands on national health budgets, but also in terms of personal suffering[9]. Despite this, few countries have surveillance systems that generate reliable information on the nature and extent of injuries, especially with regards to non-fatal injuries.
The traditional view of injuries as "accidents" or random events has resulted in the historical neglect of this area of public health[10]. Yet, for the last decades, public health officials have been recognizing injuries as preventable events and have been promoting evidence-based interventions for the prevention of injuries, worldwide[5].


Figure I.2 Public health approach to injury prevention.
Redesigned from the National Center for Injury Prevention and Control's graphic.

Many injuries can be prevented through effective approaches as seen on Figure *I.2*. For example, if injury indices are accurately calculated and reported, they can be used to develop policy and mitigate the burden of injury, as well as to optimize service provision[11].
Many injury interventions are already in place (e.g., transportation requirements such as setting speed limits, safe automobile design, seatbelt and other safety restraint use, helmet and protective equipment use, workplace safety program implementation, and ergonomic design) and achieved significant public health improvements including reduction of injury occurrence[12]. The collection, centralization and analysis of injury data depends on the resources and political will of each country, and some targeted solutions already exist in France, such as surveillance of road traffic accidents or suicide. A more comprehensive trauma surveillance system could be added to the existing French public health surveillance system. In the next section, we will describe in detail the existing trauma and public health surveillance systems in France, the requirements and the expected benefits of a comprehensive trauma surveillance system.

## I.3 Building a National Trauma Surveillance System

### I.3.1 Existing trauma related surveillance systems in France

Certain types of trauma are already the subject of targeted surveillance, two of which are road traffic accident surveillance and the suicide observatory.

#### *I.3.1.1. Road Traffic Accident Surveillance*

The purpose of Road Traffic Accident (RTA) surveillance is to examine accidents based on the severity of the injuries in order to identify ways to intervene and prevent them. This surveillance also keeps track of changes in the number of serious injuries for different types of users and areas.

In France, data on RTA are collected by the police or law enforcement authorities at the location of the accident. This data includes details such as the type and number of vehicles involved, the traffic environment, and sometimes the cause of the accident. However, there is a problem of underreporting as the police may not be informed of all types of accidents. In France, the number of injuries reported in the national road accident database by law enforcement officials is underestimated because they are not always called to the scene of accidents where there are no fatalities, especially for micro-mobility vehicles such as bikes, electric scooters, and skateboards.

Additionally, the police often rely on immediate evaluations to determine the severity of injuries recorded in road safety databases. However, misreporting is common due to the fact that the police are not equipped to conduct thorough medical assessments to determine the actual severity of an injury.

Thanks to a model built by comparing the data from the Rhône Register with that of the national file of road accidents, and by projecting at the national level the registered accidents, the Gustave Eiffel University makes it possible to estimate the number of road injuries in France. The Rhône Register has aimed since 1995 to list all the victims of road accidents that have occurred in the Rhône department. The data comes directly from 245 hospital emergency departments, whether the injured persons are hospitalized, or only treated in the ED. The Rhône Register has been able to estimate the under-reporting ratios between law enforcement reports and the data from the register and has shown great discrepancies for slightly to moderately injured persons. As an example, people aged from 14 to 19 years old who had an accident while riding a bike were 17 times more frequent within the Rhône register than in the law enforcement database from 2012 to 2016[13]. A table with all ratios between total number of law enforcement injuries and Gustave Eiffel University estimate on average over 2012-2016 is presented in Appendix H.

The implementation of the Electronic Health Record (EHR) has improved the completeness and quality of the Register data collection, yet it still requires a heavy investment of time and labor for healthcare professionals. In order to obtain a comprehensive data collection for the whole French territory, other solutions must be considered.

### *I.3.1.2. The National Suicide Observatory*

In mainland France, attempted suicide results in almost 100,000 hospital admissions and around 200,000 ED visits per year, or around 20 suicide attempts for every death[14].
The suicidal risk is higher for people with a history of suicide attempts:

- 75% of recidivism occurs within 6 months of a suicide attempt;
- The occurrence of an attempt multiplies by 20 the risk of another attempt in the following year, and by 4 the risk of a subsequent suicide.

Created in 2013, the French National Suicide Observatory (NSO) is responsible for:

- coordinating the various data producers;
- identifying research topics, prioritizing them, and promoting them among researchers;

and defining indicators for monitoring suicide prevention policy.
and its scope includes suicidal behavior as a whole, "ranging from suicidal ideation to suicide planning, suicide attempt and suicide."[15].
The NSO led to the creation of the VigilanS system[16]. Launched in 2015 in the Hauts-de-France region, the overall objective of the VigilanS system is to help reduce the number of suicides and repeat suicide attempts. The system consists of a contact and alert system that organizes a network of health professionals around people who have attempted suicide, who keep in touch with them. As of February 2023, VigilanS has been implemented in 17 regions, including 4 overseas regions, and 92 departments. Anyone hospitalized for a suicide attempt is offered enrollment in VigilanS upon discharge. At the same time, the patient's general practitioner and psychiatrist, if any, receive a letter informing them of the organization and their patient's enrollment. They also have a dedicated phone number to answer any questions they may have.
Plancke et al. showed a 15% reduction in suicide recurrence with the VigilanS program, but pointed out that patients who had attempted suicide and were not hospitalized could not benefit from this program[17].

However, the NSO has not yet reached its full potential due to several issues that need to be addressed:
The actual data sources allow identification of only suicide attempts that resulted in contact with the health care system. Indeed, Attempted suicide data on which the ONS relies are derived from the private and public hospital billing system as well as ED visit summaries that are sent to the Oscour network (detailed in section I.3.2), therefore they do not include suicide attempts that did not require hospital treatment (i.e. the least serious from a somatic point of view). Attempted suicide and suicide underestimation has been pointed by the WHO[15] and in France, the CépiDc-Inserm evaluates this underestimation at about 10%[18].

The actual data lacks detailed information about attempted suicide. Indeed, the ED visit summaries sent to Oscour comprise solely ICD-10 main and secondary diagnosis code. Previous attempts, proximal risk factors (i.e. conflict with partner or family, death of a relative, financial problems) and the suicide specific method used are absent features.

In conclusion, these national observatories rely on partial information and underestimation is a major drawback, despite the will to monitor RTA and suicide attempts.

### I.3.2 French Injury Surveillance System: a political will and an ongoing project

The benefits of injury surveillance systems are the possibility to:

- investigate: to confirm the cluster of injury issues, to get primary evidence about how and why certain injuries occur in specific risk groups;
- perform epidemiological studies;
- design and apply appropriate interventions;
- monitor the results and assess the impacts of interventions;
- provide arguments to support budget requests or resources.

#### *I.3.2.1. Benefits for road safety monitoring and prevention*

In the field of road safety, the existing French system is based on the reports produced by law enforcement. It has been shown that this system largely underestimates road morbidity (up to 70% for certain types of accidents such as bicycle accidents)[18]. Led by the Ministry of the Interior, this system has no strong link with the health system. In addition, the monitoring of the risk of RTAs related to medical incapacity to drive on the one hand and to the use of medication on the other is based on the CESIR observatory (matching law enforcement reports and medical insurance databases) which covers 20% of the drivers involved in an accident that has been reported by the police[19]. A full surveillance system would provide a more statistically powerful tool for the evaluation of the health consequences and costs of road insecurity.

#### *I.3.2.2. Drug consumption study*

The traumatic risk related to drug use does not only concern traffic accidents. Each year, more than 5 million visits to the ED are motivated by a non-road accident trauma. The French epidemiological surveillance system does not currently allow the study of this risk.
On the side of voluntary trauma, a surveillance system would constitute a powerful source of information for the study of the suicidal risk related to the consumption of drugs (e.g. isotretinoin).
The impact of accidental events on drug consumption on the one hand, and on the entry into addiction to psychotropic drugs on the other hand, are subjects that have not been explored and that could benefit fully from such an information system.

#### *I.3.2.3. Benefits for fall monitoring and prevention*

As seen in section I.2.1, falls represent the main trauma cause, and the elderly are at risk for this type of trauma[20]. Drugs have been found to be on the main risk factors for falls among this particular population[21], however, in France, the study of the association between drugs and falls is mostly based on cohorts and is not exhaustive[22]. A linkage between a surveillance system and reimbursement data would provide powerful studies for this public health problem.

#### *I.3.2.4. Benefits for violence monitoring and prevention*

In France, crime statistics are collected through law enforcement using complaint files. However, this method of collection suffers from several biases that are difficult to control (for example influenced by police forces activity levels) and makes temporal and geographical comparisons uncertain. A health services-based system for recording violence would provide an unprecedented tool for guiding public policy in terms of the fight against delinquency and domestic violence.
Moreover, in the case of domestic violence, in addition to the problem of under-representation due to the methods of collection, the current preventive measures are primary and not targeted or secondary. A detailed and precise registration in terms of location would make it possible to mobilize appropriate and targeted resources. A real-time monitoring of domestic violence would also speed up immediate political decisions to provide primary prevention and reinforce common measures such as text messages, code words given to trained people (e.g. pharmacists, shopkeepers, bar owners), the increase of shelters[23].

#### *I.3.2.5. Benefits for the French heath system*

In the area of planning and optimization of health services, an exhaustive and real-time surveillance can produce useful statistics for the entire health system. Decision-makers and healthcare managers need predictive data to adjust resources within the emergency department and downstream (hospitalization, medical and psychosocial care, prevention). For example, when several EDs can receive trauma patients in urban areas, a poorly managed allocation of resources can lead not only to a loss of chance for the patient, but also to overcrowding in a single department. As an example, the distribution of destinations for trauma victims upstream of the ED when, in urban areas, several EDs are available, is based on poorly controlled logics and is detrimental to the regional organization and planning of rescue and emergency services.

An optimal emergency surveillance system can allow a better knowledge of the distribution of diseases of the populations attending the ED and can also be a tool for the evaluation of primary health interventions. It can identify groups and neighborhoods with high rates of ED visits and measure the impact on the weeks following discharge from the ED. It can also facilitate health prevention initiatives delivered in the ED. The benefits of such a trauma surveillance system are numerous, yet, mandatory requirements have to be drawn for its optimization.

### I.3.3 Injury surveillance system requirements

Injury surveillance is defined as: "*...the ongoing systematic collection, analysis, and interpretation of injury data, for use in planning, implementation and evaluation of prevention activities. Injury prevention programs use surveillance data to assess the need for new policies or programs and to evaluate the effectiveness of those that already exist*."[24]

Surveillance produces data describing:

- the size of and characteristics of a health problem
- the population at risk
- the risk factors
- the trends

A performant and optimal surveillance system includes specific attributes such as:

- Secure and privacy-enhanced.
- Simplicity: All needed data should be produced but in the most straightforward and simple way possible. Data format should be easy to understand and complete and more importantly should not add workload or waste staff time.
- Flexibility: The system should be easy to change, especially when ongoing evaluation shows that change is necessary.
- Acceptability: Involving staff in the design, evaluation and improvement of data entries may help ensure that end users are getting the results they need from the system.
- Reliability: The system should fully record injury events will all relevant information being described and classified accorded to stated definitions. The system should also exclude non-injury events. Sampling should be avoided.
- Utility: The system should be practical and affordable. It should not put unnecessary burdens on an agency's staff and budget.
- Timeliness: The system should be able to generate up-to-date information whenever that information is needed.

As suggested by the WHO, the Minimum Data Set (MDS) for injury surveillance system includes demographic and injury-related variables, such as place of occurrence, activity, injury mechanism, intent and nature of injury[5]. The core Optional Data Set (ODS) recommended by the WHO includes variables such as the external cause of the injury, whether alcohol or another substance was a factor, the severity of the injury, the disposition of the person. Figure I.*3* illustrates the building blocks (datasets) of an injury surveillance system. An example of form is provided in Appendix I.

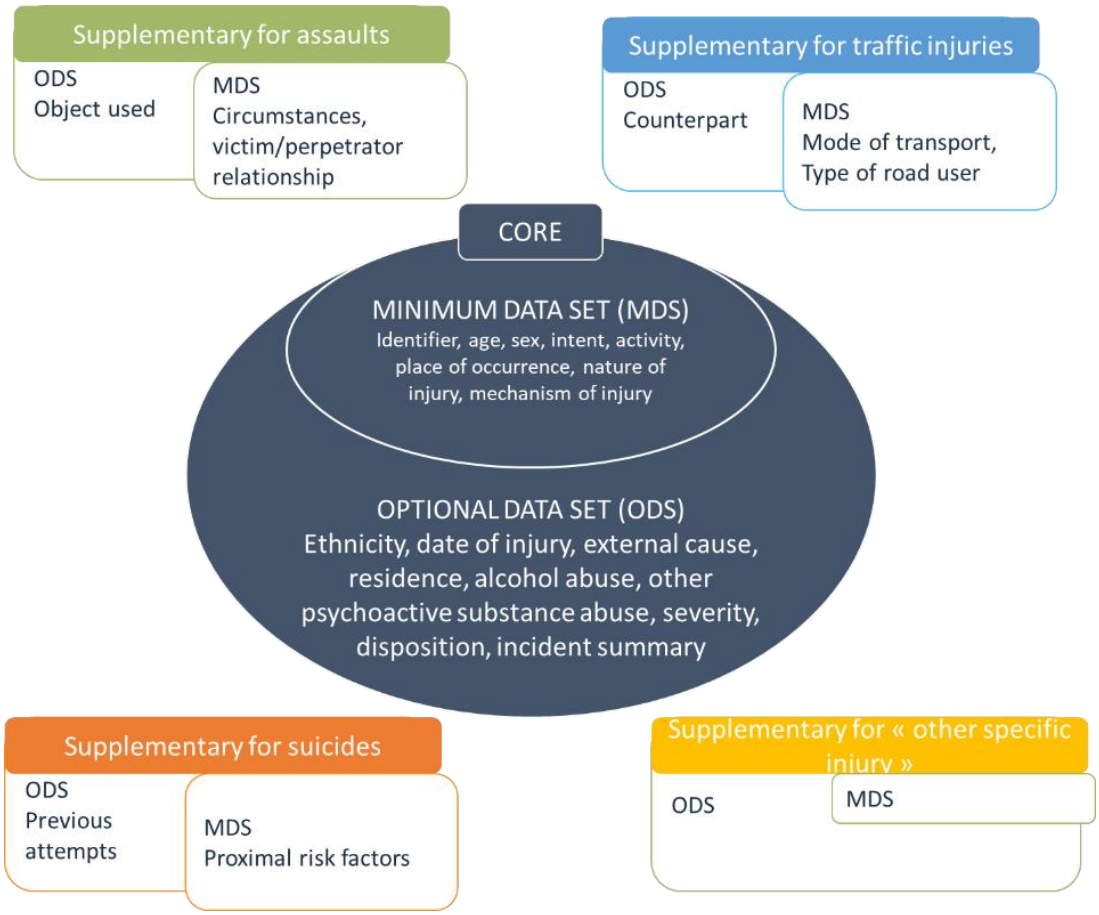



Figure I.3 Building blocks of an Injury Surveillance System

To achieve this goal and apply WHO's injury surveillance guidelines[5], many countries use digital data for injury data management[25,26]. However, the European association for injury prevention and safety promotion noted that the management of injury data is a difficult and challenging process[27].

### I.3.4 French emergency surveillance system

The lack of unified injury data collection method and the absence of a framework for cooperation among healthcare facilities are the main challenge ahead of effective injury data collection. Several data sources can provide the necessary information for an injury surveillance system: health clinic records, general practitioners' (GP) records, ED records, ICU (Intensive Care Unit) admission records, death certificates, ambulance or Emergency Medical Technician (EMT) records, police traffic accident reports, police reports… Ideally, use should be made of existing data sources and each source of data has its own set of advantages and drawbacks. For instance, ED data can be relatively complete but less reliable if healthcare providers have to spend extra time on documentation. Using existing data sources and systems provide sustainability for an injury surveillance system and the automatic and daily reporting of ED visit summaries and emergency care activities (ORU and Oscour) provides a powerful real-time monitoring tool for ED data in France as seen on Figure I.*4* . The Oscour Network was set up in 2004 by the INVS (Institut National de Veille Sanitaire), back by Santé Publique France after a heat wave with exceptional health consequences hit France[28]. The health phenomenon led, on the one hand, to massive recourse to the emergency care system

and, on the other hand, to a sudden increase in morbidity. The first perceived consequence was the saturation of the health care system, while the health monitoring services and networks, which did not have warning indicators, did not have the expected responsiveness. This crisis showed that such phenomena, in their origin, geographical scope and consequences, exist and would inevitably recur, which has been the case with the Coronavirus outbreak. The Oscour Network aims at identifying health situations requiring an adapted public health response, but also to ensure the measurement of the impact of epidemics or expected events. In fact, data is collected by directly extracting information from the patient's EHR, which is created for each patient coming to an ED. For data homogeneity purposes, a single data format called RPU (Résumé de Passage aux Urgences) has been defined. It contains several types of variables: socio-demographic, medical and hospital trajectory variables as shown in Appendix *A* Résumé de Passage aux Urgences v2, variable definition and formatAppendix A and Appendix B-F.

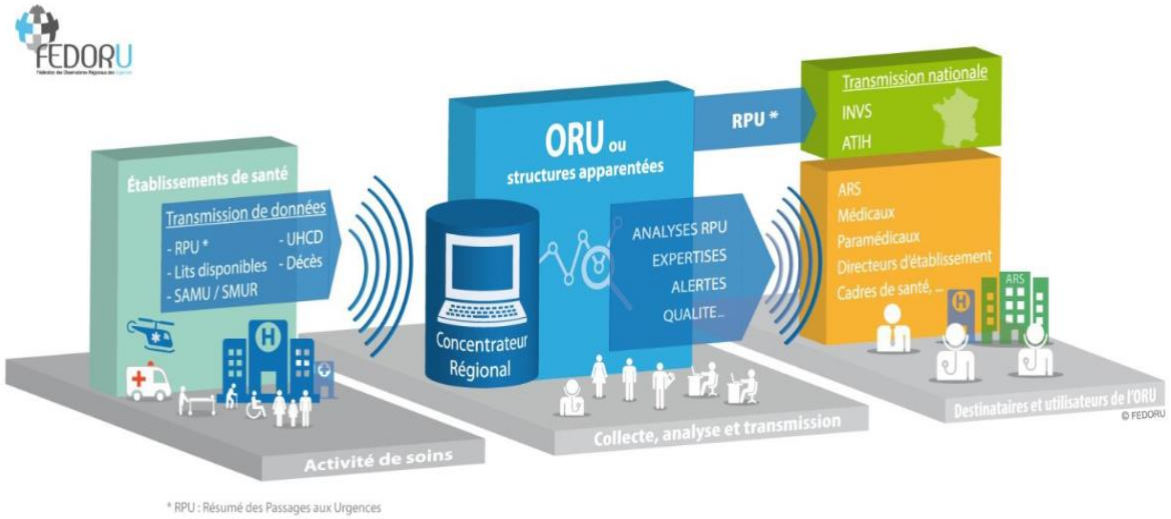



Figure I.4 Stakeholders involved in French Emergency Surveillance.
*RPU: Résumé de Passage aux Urgences, UHCD : Unité d'Hospitalisation de Courte Durée, ORU : Observatoire Régional des Urgences, INVS : Institut National de Veille Sanitaire, ATIH : Agence Technique de l'Information sur l'Hospitalisation, ARS : Agence Régionale de Santé.*

Since its inception, the Oscour network has recorded more than 130 million ED visits. On April 16, 2019, the French Health Minister, Agnès Buzyn, announced that the Oscour database would be hosted by of the Health Data Hub[29] (HDH). This Hub aims to cross-reference all French health databases and facilitate their use by research and development teams. This link between the Oscour network and other databases such as those of the social security system will make it possible to form cohorts to investigate a wide range of questions to improve health surveillance in France. In May 2022, the Oscour database was fully integrated in the HDH[30].

When considering a comprehensive and real time trauma surveillance system, the use of such a network would represent an opportunity. However, when injuries are reported, the sole information available are the main and secondary diagnoses (ICD-10 codes) and is not sufficient to establish surveillance indicators related to trauma because the type of event and the injury mechanism remain undocumented.

### I.3.5 From Electronic Health Record to ED visit summaries for trauma

The addition of trauma related MDS and ODS, as advocated by the WHO guidelines[5] to the ED visit summaries managed by the OSCOUR network would allow for an optimal near real-time injury surveillance system. If we consider this option for this surveillance system, the requirements as proposed by WHO guidelines must be assessed (section I.3.3).

#### *I.3.5.1. Mandatory structured form for trauma data*

The major constraint associated with the RPU is that the data implemented must be structured.
One could consider adding directly unstructured data to the RPU and extracting trauma related information from text once the RPU has navigated through the network and has arrived to INVS or Santé Publique France. However, clinical notes contain identifying data[31] and according to GDPR, this type of data should remain locally. Therefore, the Secure and privacy-enhanced criteria of the surveillance system would not be met.

#### *I.3.5.2. Trauma classification*

To fulfill the reliability criteria, the system should fully record trauma events will all relevant information being described and classified accorded to stated definitions. Should a specific thesaurus be used or created, or can other general classification tools or ontologies be used? ICD-10[3] and ontologies such as SNOMED[32], UMLS[33] or the Thésaurus of the French Society of Emergency Medicine (Société Française de Médecine d'Urgence, SFMU)[34] are unable to provide a detailed description of trauma mechanisms. Therefore, a specific ontology or a classification grid needs to be used or created specifically for these criteria. The International Classification of External Causes of Injury (ICECI)[35] seemed be the best option for benefiting a standardized classification tool in line with WHO recommendation. Released in 2001 (with a last update in 2004), ICECI was a system of classification designed to enable systematic description of how injuries occur. ICECI had a multi-axial and hierarchical structure: core module including seven items (mechanism of injury, objects/substances producing injury, place of occurrence, activity when injured, the role of human intent, use of alcohol, use of (other) psycho-active drugs) and five additional modules to enable the collection of additional data on special topics (violence, transport, place, sports, occupational injury). However, no French version of the ICECI was released, validated and tested. And, as mentioned in the Discussion section above, in 2022, the WHO announced that the ICECI was no longer maintained. Other solutions have to be envisaged and will be detailed in the discussion part.

#### *I.3.5.3. Emergency Department EHR data*

When a patient arrives at the ED, several mandatory and optional information and data are implemented in the EHR by administrative and medical professionals, such as age, weight, blood pressure, chief complaint, mean time of arrival... As a reminder, an optimal trauma

monitoring system implies simplicity, so it should not increase workload or waste staff time, and if ED professionals had to implement all the necessary criteria for MDS and ODS, this would represent up to 15 fields, which is time consuming. The trauma-related information needs to be captured elsewhere. Much of the available clinical data is in narrative form (unstructured) often named clinical note, resulting from transcription of dictation, direct entry by providers, or the use of voice recognition applications[36,37]. A clinical narrative or note is a brief summary of specific events experienced by patients. It describes imaging observations, physical symptoms, and in the case of emergency medicine, circumstances of trauma, such as the location, activity or consumption of substances. This type of data is called "unstructured": it cannot be easily organized using pre-defined structures, unlike structured data which is organized into specific fields as part of a schema, with each field having a defined purpose.
Narrative data account for a large component of the information that is gathered in the care of patients. Studies mention a proportion of 70% of unstructured data for ED EHR[38] or 75% for general EHR[39]. Furthermore, a great proportion of data related to trauma mechanisms is available in clinical notes as seen in the example below on Figure I.5.

Chute mécanique d'une marche de caravane en état d'ébriété. TC sans PCI. Douleur de l'hallux G avec impotence fonctionnelle totale.

*Figure I.5 Example of a trauma-related clinical note extracted from Bordeaux University Hospital database. TC: Trauma Crânien, PCI: Perte de Connaissance Initiale, G: Gauche*

In conclusion, the specifications for adding defined trauma information to the French emergency surveillance system are to extract information from unstructured clinical notes written in French without adding workload for healthcare professionals. Extracting trauma information from unstructured text data without human intervention can be performed with Natural Language Processing (NLP) and this specific area has been the subject of numerous research previously. The following chapter will navigate through this specific matter.

# II. NATURAL LANGUAGE PROCESSING FOR CLINICAL DATA

The rapid adoption of EHR and the parallel increase in narrative data in electronic form, along with the need to improve the quality of care and reduce medical errors are both strong incentives for the development of Natural Language Processing (NLP). Much of the available clinical data is in narrative form as mentioned above[36,37]. This free-text form is convenient for expressing concepts and events, but is difficult for research, summarization, classification, decision support or statistical analysis. To reduce errors and improve control, labeled data is needed. This is where text mining and NLP, specifically information extraction (IE) and classification, is needed.

## II.1 Narrative Clinical data

Narrative data account for a large component of the information that is gathered in the care of patients. A narrative tells a story: it allows seeing the patient through a description and complicated events are easier to describe in text rather than filling them in codes. It has a lot of contexts and can be both alphabetical and numerical data. Clinical notes contain objective and subjective assessments of a patient's condition. Such raw notes contain the intuitions and observations healthcare professionals who regularly monitor the patient. This valuable patient-specific information present in clinical notes has the potential to uncover hidden clues about the mental state and the health of a patient[40]. Several challenges come along with the analysis and modeling if clinical notes such as high-dimensionality, rawness, sparsity, and linguistic complexity[41]. Healthcare professionals prefer using natural language and free text for documentation over restrictive structured forms[42], but healthcare professionals have adapted to time-intensive note-writing by relying on overloaded and inconsistent medical and language abbreviations as well as rich medical jargon[43].

Some issues come along with the unstructured form of clinical notes:

1. Differences of length at various scales:

    Patients: for patients who may have had complex circumstances or multiple histories with numerous encounters with the health care system, clinical note can be dramatically long[44].

    Healthcare providers: some of the professionals are more inclined to write exhaustively than others[45]. Some professionals might also find it easier to type on a computer while using more than two fingers for typing[46,47]. Furthermore, using computers while taking care of a patients require multi-tasking skills which directly depend on practitioner's baseline skills[48,49].

    Time: writing clinical notes is time consuming, therefore their length can vary depending on the moment of the day/week/month and the workload of the ED department[50].
2. Heterogeneity of language at different levels:

Healthcare providers: disparities can be observed within and between health care professionals[45] from a gender[51], social groups[52] or individual personalities[53] perspective. Differences could also be seen between graduated professionals and students.

Hospital: language, abbreviations and even the type of information collected leading to a clinical note can vary between ED departments.

Region: language, abbreviations and expressions are different from a region to another.

Narrative information, when expressed in many ways, can be ambiguous.

*Textbox II.1 Narrative information ambiguity*

*Ambiguity* as a prominent obstacle in the field of NLP. When attempting to comprehend the intended sense of a word, multiple factors come into play. These factors encompass the contextual usage of the word, our personal understanding of the world, and the conventional usage of the word within society. The meanings of words are subject to change over time, and they can also exhibit divergent interpretations across various domains. This phenomenon becomes evident in instances of homographs, where two words share identical spellings but originate from different etymologies. Furthermore, polysemy exemplifies the occurrence of a single word carrying multiple distinct meanings.

### II.1.1 Natural Language and Sub-languages

Sublanguage, a subset of natural language, is another challenge for NLP. Medical language is a sublanguage with a subset of vocabulary and different vocabulary rules from the main language. To extract meaning from sublanguage, NLP systems must understand the rules of that language. Social media, for example, is a sublanguage. It uses abbreviations and emoticons to express meaning (versus using words for the same concepts). With these differences, analysts cannot run an NLP system trained on newspaper text on social media and expect it to extract the meaning.

Medical language has different sublanguages within it. For example, medical blogs and clinical notes use different language. Because of these differences, health systems should not purchase or use off-the-shelf NLP systems built for one sublanguage and use it on another. Developers and analysts must tailor NLP systems for use on a specific language (e.g., healthcare)., and that tailoring process takes time. Patient comments such as "I'm dizzy" or "my stomach hurts" can tell clinicians a lot about a person's health, as can other information such as zip code, employment status, access to transportation, and so on. This critical information, however, is captured as free text, or unstructured data, making it impossible for traditional analytics tools to exploit.

## II.1.2 Natural Language Processing for French Clinical Textual Data

In addition to the semantic peculiarities due to the specific nature of clinical notes, the language used is a non-negligible component. The number of published articles related to NLP in medicine had increased drastically for the last two decades and France is one of the most productive countries as shown by Wang et al.[54]. In 2018, Névéol et al. counted 111 publications when searching for clinical NLP in other languages than English[55].

### *II.1.2.1. Natural Language Processing tasks*

When processing clinical data from EHR, 4 main tasks can be defined:

- Text classification
- Clinical Named Entity Recognition[1]
- Relation extraction
- Others (i.e. information retrieval, natural language generation, abbreviation disambiguation.)

In a systematic review conducted by Wu et al. in 2020, it was found that text classification using deep learning on clinical data from HER was the most prevalent task in the literature, accounting for 40.5% of the studies, followed by Named Entity Recognition (34%) and relation extraction (13.7%)[56].

They also pointed out that the top clinical tasks were (as seen on Figure II.1):

- Clinical concept extraction (i.e., the extraction of common clinical concepts, such as problem, lab test, treatment, time expressions, and events)
- Phenotyping (i.e., the broad characterization of patients' conditions)
- Clinical relation extraction (i.e., the identification of relations between the common clinical concepts)

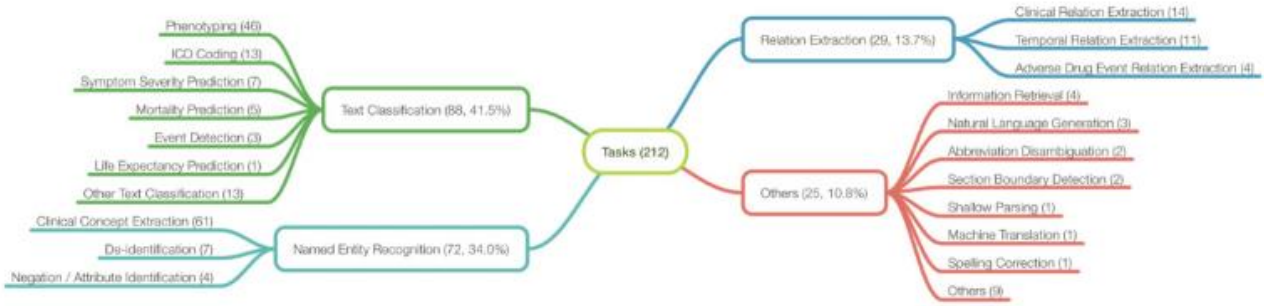



*Figure II.1* Tasks and their prevalence in Wu et al. study from the NLP and the clinical perspective.
Most journal articles analyzed comprised several tasks, hence the percentages[56].

In order to obtain the necessary information for the implementation of the trauma surveillance system several strategies were considered.

[1] Clinical Named Entity Recognition (NER) is a critical natural language processing (NLP) task to extract important concepts (named entities) from clinical narratives.

For example, for the 2018 CLEF eHealth Task 1 challenge[57] which objective is to extract ICD-10 codes from death certificates provided by the CépiDc (Centre for Epidemiology of Medical Causes of Death), several strategies were considered. The team of Cossin et al.[58] tested an approach based on ontologies, while Flicoteaux et al.[59] proposed an approach using a probabilistic CNN (Convolutional Neural Network) and Amin-Nejah and al.[60] resorted to the association of a RNN with a CNN.

For our project, Named Entity Recognition was envisaged due to the large possibilities of using ontologies and terminologies. Indeed, French medical NLP can benefit from initiatives such as HeTOP[61] for mapping onto-terminology codes. This initiative held by the Rouen Hospital gather more than 3 millions of concepts with 100 terminologies and ontologies. Using onto-terminologies such as UMLS, SNOMED-CT or the thesaurus of the Société Française de Médecine d'Urgence (as mentioned in section I.3.5.2) with NER could have been an optimal solution. However, many necessary concepts such as the type of suicide or the type of counterpart in RTA (Road Traffic Accident) are not available in French onto-terminologies.

As an example, Metzger et al., while classifying French ED clinical notes for suicide, extracted medical and clinical concepts with UrgIndex which is a tool for extracting and encoding medical concepts from ED clinical notes[62]. This method has shown good performance on intra-hospital syndromic surveillance with an overall recall of 85.8% (95% CI: 84.1-87.3) for respiratory syndromes and cutaneous[63]. However, despite the use of a powerful French-language medical multi-terminology indexer developed by the CISMeF (Catalogue and index of French-language medical sites)[64] which comprises several termino-ontologies, this type of tool cannot be used for retrieving a precise description of trauma mechanism.

Without a validated French trauma classification tool (i.e. ICECI), mapping is currently not possible. Hence, the chosen task for our project was classification, and the rest of this manuscript will cover this particular area.

#### *II.1.2.2. Related Work Summary*

As seen on Figure II.1, the main goals of classification for clinical NLP are phenotyping, ICD-10 coding, and symptom severity description. Considerable work has been done on text classification for medical texts and clinical notes. These works are mostly based on traditional and classical methods such as support vector machines (SVMs), Naives Bayes, random forest or k-nearest neighbor (kNN). Recently, deep learning has gained great attention from researchers for addressing classification tasks on clinical notes as seen on Figure II.2[56].

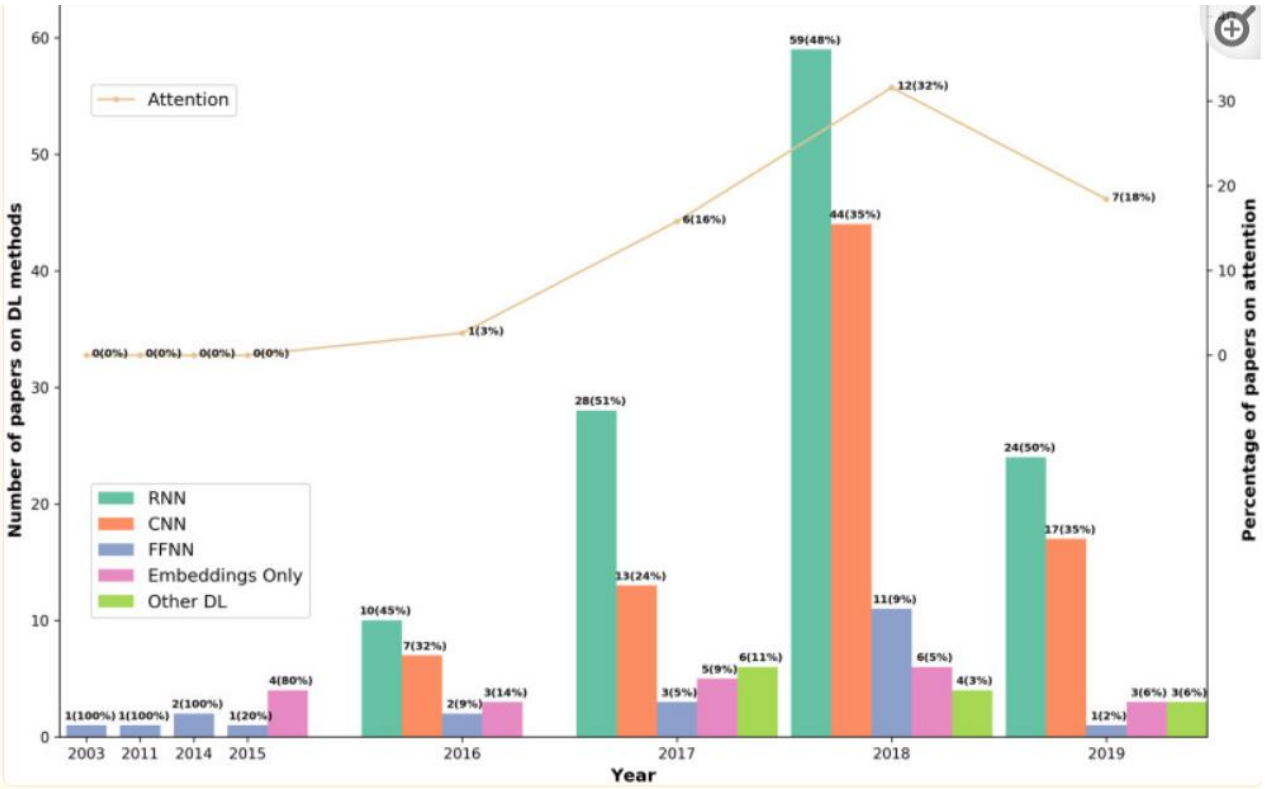



*Figure II.2 Growth of broad architectures in deep learning over the years. Wu et al.*[56]

In 2020, Wu et al. showed in their review that RNN variants were the most common (n = 129, 60.8%) deep learning methods used for all NLP tasks concerning EHR, CNNs were second (n = 83, 39.2%), traditional feed-forward networks were third (n = 22. 10.4%), and embeddings-only were fourth (n = 21, 9.9%)[56]. For text classification, in recent years, CNNs attracted attention and achieve competitive results on classification tasks for EHR[65]. Meanwhile, models using attention mechanism have gained increasing interest. NLP has seen a recent breakthrough with the introduction of deep learning, and particularly the transformer architecture. Introduced in 2017 by Google and proposed in the article Attention is All You Need by Vaswani et al. [66], transformers have an architecture that allows the implementation of a mechanism for processing the sequence of tokens that form a sentence in a self-attentive way, i.e. relating each of these tokens to each of the others in the sentence. They have the particularity of being able to be pre-trained from a corpus of text which can be very large since it does not require a coding stage. This phase leads to a generative model which is capable, for example, of constructing artificial text by iteration. The Bidirectional Encoder Representations from Transformers (BERT) is one of these transformer-type models pre-trained on large corpora of text[67]. The BERT model is a bidirectional transformer, composed only of encoder blocks. Bidirectional indicates that BERT learns information from both the right and the left side of a token's context during the (pre)-training phase. BERT is composed of a stack of N =12 identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. In other words, text encoder turns text into a numeric representation. For many tasks, including text classification, its performance is systematically superior to the convolutional and auto-regressive models used until then[67].

French derivatives such as FlauBERT[68] and CamemBERT[69] of the BERT model have been trained on very large and diverse French corpora. FlauBERT is a French BERT trained on a very large

and heterogeneous French corpus. Models of different sizes were trained using the Jean Zay supercomputer of the CNRS (Centre national de la recherche scientifique). There are three sizes: small (54 million parameters), base cased (138M) and uncased (137M) as well as large (373M). CamemBERT is based on RoBERTa[70], an evolution of BERT in several aspects, including the use of the masked language model as the sole pre-training objective. CamemBERT, like FlauBERT, is available with different sizes: base (110M) and large (335M), but also with different training corpora such as OSCAR (either 138GB or 4GB of text)[71], CCNET (either 135GB or 4GB)[72] or French Wikipedia (4GB).

One of the most interesting examples of transformer architecture is GPT-2, released by OpenAI in 2019. GPT-2 (Generative Pre-Training 2) is a large transformer-based model, composed solely of decoder blocks, with 1.5 billion parameters on its extra-large version, trained on a dataset of 8 million web pages to predict the next word from the previous words.17 Three other sizes of GPT-2 have been released before the largest one: with 124 (small), 355 (medium), 774 (large) million parameters. This model's ability for text generation quickly attracted the attention of the community because of the difficulty to distinguish the artificial texts produced from texts written by humans, suggesting that some of the meaning present in natural language was embedded. Moreover, beyond its ability to generate coherent texts, the GPT-2 can perform other tasks such as answering questions or classifying documents. As with BERT, the conservation of several self-attention blocks weights from a pre-trained model is sufficient to transfer contextual representations into another dataset. The training of the GPT-2 model is thus carried out in two distinct phases: the first phase of self-supervised generative pre-training, consists of the reading of a corpus of texts. It leads to the ability to generate texts automatically. The second supervised training phase consists in resuming the learning process from a corpus of annotated texts in order to create a system capable of performing specific tasks (classification for example). BelGPT2 is a Belgian small GPT2 pre-trained on French corpus of 60GB (Common Crawl, Project Gutenberg, Wikipedia, EuroPARL...) that was released at the end of 2020[73]. Recent studies have shown the effectiveness of transformers on classification tasks for EHR free-text data such as ICD coding[74,75], phenotyping[76] or readmission prediction[77].

We propose, in the next section, to go deeper in the NLP pipeline as well as models' architecture and performances for EHR clinical text classification and evaluation.

## II.2 Natural Language Processing for Text Classification

### II.2.1 Setting the Frame

#### *II.2.1.1. Descriptive versus Predictive tasks*

To apply text mining algorithms or NLP models to medical and/or clinical data, an understanding of the nature of data mining algorithms and their functions is useful. Descriptive and Predictive algorithms are two categories of data mining or NLP algorithms[78,79] as seen on Figure II.3:

- Descriptive: data mining groups data by determining the object's similarity (or clinical notes) and detecting patterns that are unknown, or associations in the data whereby users are unable to recognize in a massive data pool. Descriptive data is investigative.
- Predictive: data mining that comprises classification, regression, time series analysis and prediction[80] implies predicting rules from training data then the rules are employed to unpredicted/unclassified data[81].

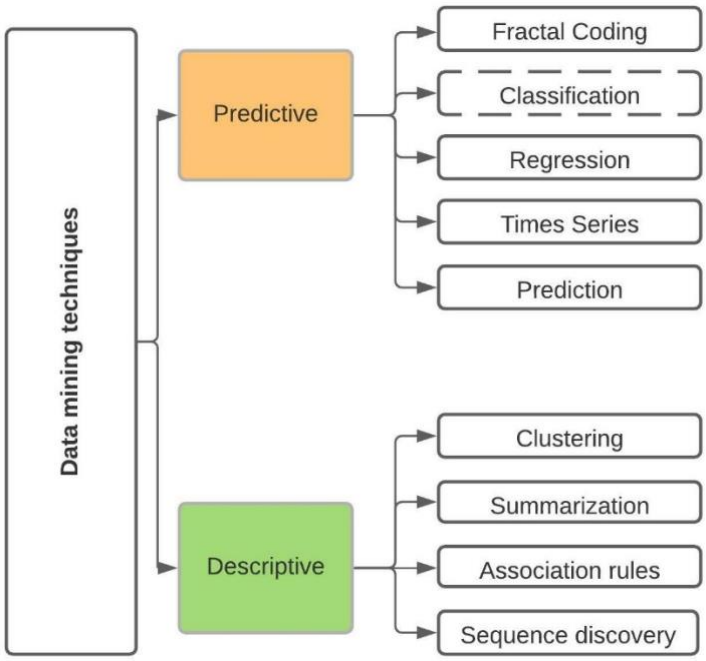



*Figure II.3 Data Mining Techniques*

On one hand, descriptive mining tasks provide the general data properties in the database and provides characteristics and descriptions without having any predefined target. On the other hand, for predictive mining tasks, inference is made on explicit values based on patterns identified by known results.

### *II.2.1.2. Learning Approaches of Text Mining and NLP*

The 3 learning approaches in data mining and NLP algorithms are:

- Supervised: the algorithm works with a set of examples with known labels whose values are nominal/categorical in classification task, or numerical in regression task.
- Unsupervised: the algorithm aims at grouping examples according to the similarity of their attribute values with unknown labels, characterizing a clustering task.
- Semi-supervised: learning is conducted when there is availability of a small subset of labelled examples, concurrent with many unlabeled examples (i.e. few-shot learning).

The task of categorizing is regarded as a supervised or semi-supervised technique in which each instance belongs to a given class, specified by the value of a special goal attribute or the class attribute[82].

#### *II.2.1.3. Text Mining and NLP differences*

In addition, a delineation must be set clear between Text Mining and NLP. Both Text Mining and NLP aim to extract information from unstructured data. Text mining is concentrated on text documents and mostly depends on a statistical and probabilistic model to derive a representation of documents. NLP tries to get semantic meaning from all means of human natural communication like text, speech or even an image.

The term text mining is used for automated machine learning and statistical methods used for this purpose. It is used for extracting high-quality information from unstructured and structured text. Information could be patterned in text or matching structure but the semantics in the text is not considered.

#### *II.2.1.4. Development life cycle*

For developing an NLP system, the general development process will have the following steps:

- Understand the problem statement.
- Decide what kind of data or corpus we need to solve the problem. Data collection is a basic activity toward solving the problem.
- Analyzing collected corpus. What is the quality and quantity of the corpus? According to the quality of the data and problem statement, we need to do pre-processing.
- Once done with pre-processing, start with the process of feature engineering. Feature engineering is the most important aspect of NLP and data science-related applications. Different techniques like parsing, semantic trees are often used for this unless we know in advance which features we are looking for (i.e. classification of disease).
- Having decided on an extracted features from the raw pre-processed data, we are to decide which computational technique is used to solve our problem statement, for example, do we want to apply machine learning techniques or rule-based techniques? For modern NLP systems, advanced ML models based on Deep Neural Networks are used most of the time.
- Now, depending on what techniques we are going to use, we should read the feature files that we are going to provide as an input to your decision algorithm.
- Run the model, test it and look for best parameters.
- Iterate through the above step to get the desired accuracy.

#### *II.2.1.5. NLP Pipeline*

Several steps are carried out during NLP process. Each step of the text mining and NLP pipeline as seen on Figure II.4 depends on the task we aim to perform and not all pre-processing steps are mandatory.

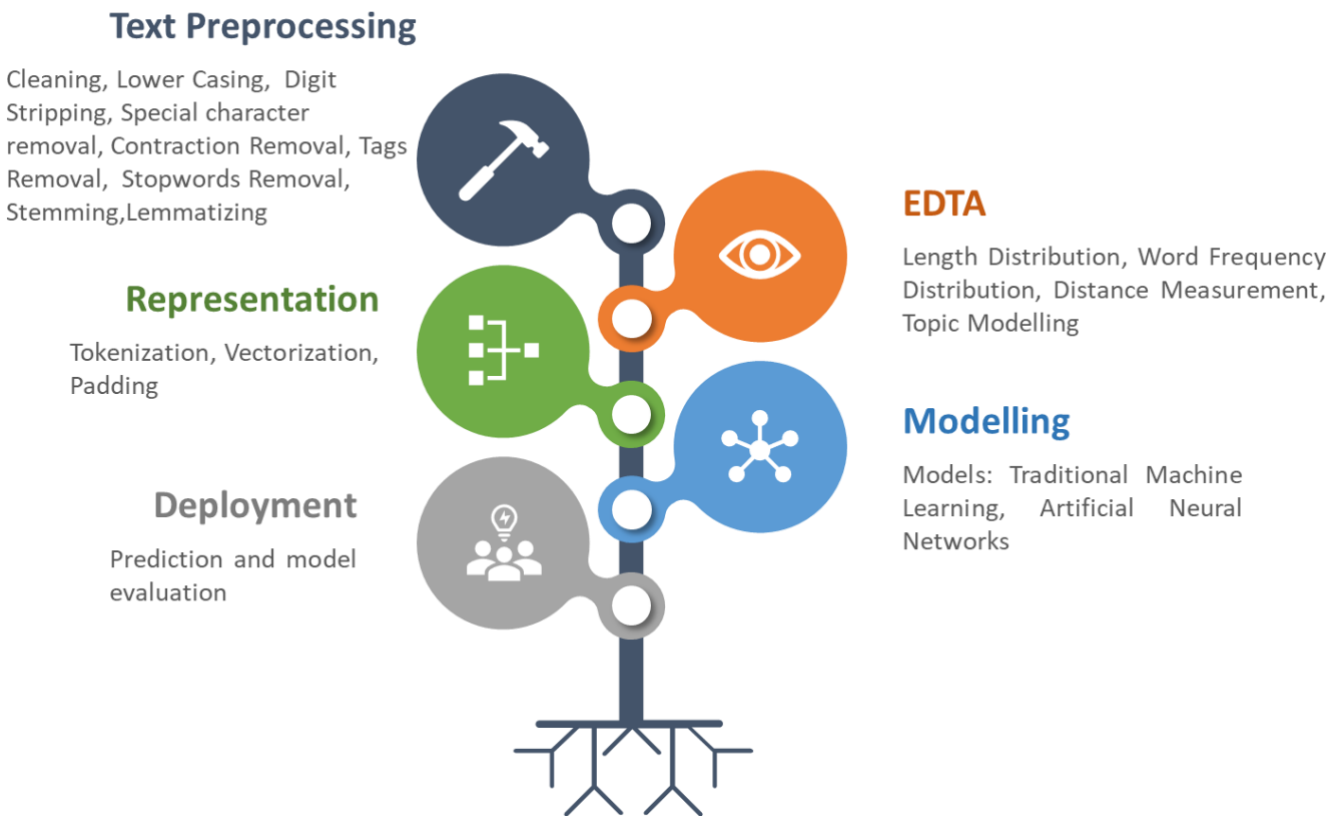



*Figure II.4 Natural Language Processing Pipeline*

*ETDA: Exploratory Text Data Analysis*

## II.2.2 Document representation and feature selection

One of the challenges associated with modeling text is the presence of inherent messiness or noise in textual data. This can pose difficulties for machine learning algorithms, which tend to perform optimally when dealing with well-defined inputs and outputs of fixed length.

### *II.2.2.1. Pre-processing*

The first step in text categorization is to normalize the text. Several techniques are performed depending on the vectorizer and the classifier model chosen such as:

- Stopwords removal: stopwords are frequent words that carry no information (i.e. pronouns, prepositions, conjunctions) in traditional machine learning
- Tags removal: URL, special characters such as "@", "https://"
- Lower casing
- Expanding contractions: in English contraction is the shortened form of a word like don’t stands for do not, aren’t stands for are not
- Punctuation removal
- Digits removal
- Extra white space removal
- Stemming: a process to reduce the word to its root stem for example run, running, runs, runed derived from the same word as run. Stemming removes the prefix or suffix from word like “ing”, “s”, es, etc. The stemming technique is not used for production

purposes because it is not so efficient technique and most of the time it stems the unwanted words.

- Lemmatizing: similar to stemming, used to stem the words into root word but differs in working. In fact, Lemmatization is a systematic way to reduce the words into their lemma by matching them with a language dictionary.

Not all cleaning techniques have to be applied to a given corpus and their use depend on the vectorizer/tokenizer and classifier chosen and the aim of modeling.

The pre-processing step for EHR clinical data is important, but no clear consensus process was identified for EHR clinical notes. As an example, removing negation words can modify the meaning of a given concept, specific tools should be used when vectorizers are used[83]. Speculation words such as "might", "suspected" should also be taken into account when classifying moderate symptoms. Redundancy can also impact performances when performing exploratory data analysis, hence it should be considered before this step[84]. Expanding contractions or abbreviations has also been the subject of research and has brought a lot of controversy. Indeed, abbreviations in clinical notes have several meanings and rule-based disambiguation decreases the generalizability of a model. Authors have been using the UMLS in order to reduce ambiguity. However, 31% of UMLS abbreviations have multiple meanings[85]. Regarding the word reduction step, Pomares Quimbaya et al. showed that adding a fuzzy and a stemming step improved the recall on a NER task[86].
Anonymization, even if our project does not comprise such a step, can also represent a pre-processing step for clinical notes. Information loss due to data manipulation can reduce NLP tasks. As an example, Stavros et al. concluded that, the loss of predictive power as a function of the information loss due to aggregation and suppression process varies considerably, depending on the nature of the chosen classifier (traditional machine learning) on Greek clinical notes[87].

#### *II.2.2.2. Exploratory Text Data Analysis (EDTA)*

The structure of a corpus can be assessed regarding its structure, its vocabulary or linguistic features and its topics. Exploratory text data analysis is a critical component of NLP and text mining research. It involves using statistical and visualization techniques to gain insights into textual data, identify patterns, and explore relationships between different variables. The purpose of EDTA is to uncover underlying structures, trends, and features in the data that can inform subsequent analysis and modeling. Identifying key topics and themes that are prevalent in the documents as well as the relationships between these topics and contextual variables is also an important step before modelling.

##### Length structure

One can assess the length distribution of a corpus by examining the word count of each document and computing the average length. This approach may aid in determining the maximum length and in padding for a tokenizer. To identify outliers, the median absolute deviation (MAD) can be utilized. The MAD is a reliable measure of variability similar to the standard deviation[88]. The MAD is defined as the median of the absolute deviations (see the

equation below) from the data's median which reduces the effect of extreme outliers. This is especially important in high-tail distributions.

$$MAD = median(|X_i - \tilde{X}|) \quad (II.1)$$

With: $\tilde{X} = \text{median}(X)$

The length of EHR clinical notes can vary depending on their type (e.g. progress note, discharge summary) or the department or ward from which it is written. In other departments than ED, notes are often "bloated": excessively long and often filled with information considered redundant and unnecessary for clinical decision-making. Use of copy/paste as well as shortcut features to drop in large blocks of templated text contribute to bloated notes that are hard to navigate, lack clinical value, and may contribute to safety risks and diagnostic errors[89]. In ED settings, patients mostly come once for a given medical or trauma problem, therefore, there is no use of copy/paste and redundancy may be limited. However, the length of the ED clinical notes might be influenced by the history of the patient or their age[90].
When considering using Transformers for clinical notes, the prior analysis of clinical notes' length can influence the choice of the maximum sequence length for a given model or even the type of model. Indeed, with Transformer models, there is a limit to the lengths of the sequences that can be can passed to the models. Most models handle sequences of up to 512 or 1024 tokens and will crash when asked to process longer sequences. There are two solutions to this problem:

- Use a model with a longer supported sequence length such as Longformer or
- Truncate the sequences.

## Vocabulary structure

When documents belong to given categories (if known beforehand) or when comparing two corpora vocabularies, the structure of each category or corpus can be modelled by extracting the frequency distribution of the unique words in each corpus or category (here $p_i$) and testing the similarity. To measure the distance between two corpora vocabularies or categories, there are several methods that can be used.

### *Euclidean distance:*

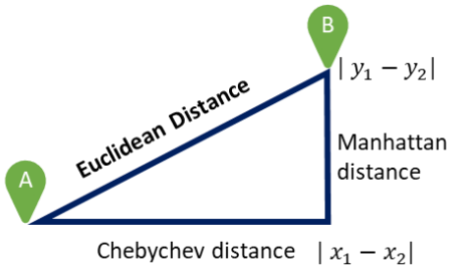



Figure II.5 Euclidean distance illustration

This method calculates the distance between two points (in this case, the frequency distribution of the words in each corpus or category) by taking the square root of the sum of the squared differences between the corresponding elements of the two vectors as seen on Equation II.2. The smaller the Euclidean distance, the more similar the two corpora are since the distance calculated using this formula represents the smallest distance between each pair of points.

$$d(\boldsymbol{P},\boldsymbol{Q}) = ||\boldsymbol{P}-\boldsymbol{Q}||_0 = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

$$= \sqrt{(p_1-q_1)^2 + ((p_2-q_2)^2 + \ldots + (p_n-q_n)^2} \qquad (II.2)$$

*Cosine similarity*

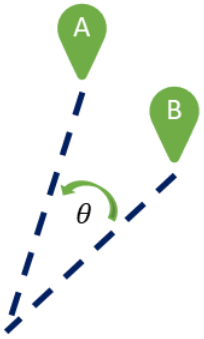



Figure II.6 Cosine similarity illustration

This method measures the similarity between two vectors by calculating the cosine of the angle between them. The closer the cosine value is to 1, the more similar the two corpora are (in terms of words frequency distribution).

$$\begin{aligned}\text{Cosine Distance} &= 1 - \text{Cosine Similarity}\\ &= 1 - \cos(\mathbf{P},\mathbf{Q})\\ &= 1 - \frac{\text{P.Q}}{||\text{P}||.||\text{Q}||}\end{aligned}$$

$$= 1 - \frac{\sum_{i=1}^{n} P_i.Q_i}{\sqrt{\sum_{i=1}^{n} {P_i}^2}.\sqrt{\sum_{i=1}^{n} {Q_i}^2}} \qquad (II.3)$$

*Pearson correlation distance*

The Correlation distance quantifies the strength of the linear, monotonic relationship between two attributes. Furthermore, It uses the covariance value as an initial computational step. However, the covariance itself is hard to interpret and doesn't show how much the data are close or far from the line representing the trend between the measurements.

$$\text{Correlation\_Distance} = 1 - \text{Correlation\_Similarity}$$
$$= 1 - \frac{Covariance(P,Q)}{\sqrt{Variance(P)}.\sqrt{Variance(Q)}}$$

$$= 1 - \frac{\sum_{i=1}^{n}(p_{ij} - \left(\frac{1}{n}\right).\sum_{j=1}^{n} p_{ij}).(q_{ij} - \left(\frac{1}{n}\right).\sum_{j=1}^{n} q_{ij})}{\sqrt{\sum_{i=1}^{n}(p_{ij} - \left(\frac{1}{n}\right).\sum_{j=1}^{n} p_{ij})^2}.\sqrt{\sum_{i=1}^{n}(q_{ij} - \left(\frac{1}{n}\right).\sum_{j=1}^{n} q_{ij})^2}} \quad (\text{II}.4)$$

Assessing the vocabulary structure, when clinical note datasets are concerned, can be performed prior to a given NLP task or afterwards when performing an error analysis. As an example, when Wang et al. compared word embeddings from several corpora, the correlation coefficient used by the authors was the Pearson one. Despite some investigations into alternatives, cosine similarity has persistently remained the default choice across research into similarity measures for textual embeddings[91]. However, Zhelezniak et al. showed that in practice, for commonly used word vectors, cosine similarity is equivalent to the Pearson correlation coefficient, motivating an alternative statistical view of word vectors. They also characterized when Pearson correlation was applied inappropriately and showed that these conditions hold for some word vectors but not others, providing a basis for deciding whether cosine similarity is a reasonable choice for measuring semantic similarity [91].

## Topic Modelling

Topic modeling is a technique in NLP that is used to discover underlying topics or themes in a collection of documents. It is an unsupervised machine learning approach that analyzes patterns of word co-occurrence in the text data to identify the main topics or themes present in the corpus.

### *Latent Dirichlet Allocation (LDA)*

The most popular topic modeling technique is Latent Dirichlet Allocation (LDA)[92] which is a generalization of Probabilistic Latent Semantic Analysis (PLSA). LDA assumes that each document is a mixture of a small number of topics, and each topic is a distribution over words. By analyzing the frequency of words in each document, LDA infers the topics present in the corpus and their distribution across documents.
Let's denote:

- $\mathbf{D} = (\mathbf{w}_1, \dots, \mathbf{w}_D)$ : a collection of documents
- $z_{d,n}$ : the chosen topic for the word $\mathbf{w_{d,n}}$
- $\theta_d$ : the topic distribution of the document $d$

- $\alpha$ and $\eta$: a priori distributions of, respectively, $\theta$ and $\beta$ where $\beta_k$ describe the distribution of topic $k$

The main goal of LDA is to determine the a posteriori distribution of the hidden variables given the document (and parameters $\alpha$ and $\beta$):

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \quad (II.5)$$

This distribution is very difficult to compute, therefore an exact inference using the a posteriori distribution of the parameters is impossible. But there are several approximate inference methods that can be used for LDA, using for example the variational approximation or the Markov chain Monte Carlo. LDA's major drawback is that it can produce ambiguous or incoherent topics, especially if the data is noisy, sparse, or heterogeneous. LDA relies on the assumption that the words in each topic are related and meaningful, but this may not always be the case. For example, some words may have multiple meanings, some topics may overlap or be too broad, and some documents may contain multiple or unrelated topics. Another disadvantage of LDA is that it can be computationally expensive and time-consuming, especially if the data is large, the number of topics is high, or the model is complex. An advantage of LDA is that it is a flexible and adaptable method, that can be applied to different types of text data. The number of topics, the hyperparameters, and the evaluation metrics can also be customized. Although LDA is
effective at modeling long conventional text collections, it performs poorly on less conventional text such as short documents[93,94]. Other algorithms like the biterm topic model (BTM)[95] are successful in modeling topics in short texts such as ED clinical notes. However, a common weakness between LDA and BTM is that the models cannot capture the context of words. As a result, the models lack generalizability across different domains, where different words are used to describe similar concepts[94].

### *BERTopic*

Recently, a new approach of Topic modeling has been proposed by Grootendorst in "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure"[96]. BERTopic is a topic model that extracts coherent topic representations through the development of a class-based variation of TF-IDF. More specifically, BERTopic generates document embedding with pre-trained transformer-based language models, clusters these embeddings, and finally, generates topic representations with the class-based TF-IDF procedure. A Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) and a Hierarchical Density-Based Spatial Clustering of Applications with Noise[97] (HDBSCAN) are used to, respectively, reduce the dimensionality of document embeddings and model clusters. Studies comparing LDA to BERTopic are yet to be conducted. However, some authors managed to show that BERTopic HDBSCAN shows superior performance with regards to topic coherence and diversity in a university-wide model[98].

Topic Modelling can also be a suitable approach to deal with the few resources available for clinical text mining in other languages than English. As an example, Lebeña et al.[99] tested LDA and Partially Labelled Latent Dirichlet Allocation (PLDA), the supervised approach of the former, for an ICD-10 classification task on Spanish clinical notes. They evaluated their methods with metrics to determine topic coherence and the relationship between topics and ICD labels; they found that PLDA using a multi-layer perceptron as a classifier had better results than with KNN. They also argue that PLDA is interpretable and aid the explainability in artificial intelligence (XAI) since PLDA promotes topic-to-ICD semantic consistency while conveying human intuition of keywords.

#### *II.2.2.3. Vectorization*

Vectorization is the general process of turning a collection of text documents into numerical feature vectors. The concept is to get some distinct features out of the text for the model to train on, by converting text to numerical vectors. A popular and simple method of feature extraction with text data is called the bag-of-words model of text.

##### Bag of Words (BOW)

A very common feature extraction procedure for sentences and documents is the bag-of-words approach (BOW). In this approach, we look at the histogram of the words within the text, i.e. considering each word count as a feature[100].
A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves:

- A vocabulary of known words
- A measure of the presence of known words.

It's the simplest of all the techniques. It involves three operations:

- Tokenization: First, the input text is tokenized. A sentence is represented as a list of its constituent words, and it’s done for all the input sentences.
- Vocabulary creation: Of all the obtained tokenized words, only unique words are selected to create the vocabulary and then sorted by alphabetical order.
- Vector creation: Finally, a sparse matrix is created for the input, out of the frequency of vocabulary words. In this sparse matrix, each row is a sentence vector whose length (the columns of the matrix) is equal to the size of the vocabulary.

The Bag-of-words model is an orderless document representation. Documents are described by word occurrences while completely ignoring the relative position information of the words in the document.
When dealing with a large corpus of text, certain words (such as "the", "a", and "is" in English) are expected to appear frequently and are therefore of little significance in conveying meaningful information about the content of the document. If the count data for these words is utilized directly in a classifier, they would dominate the frequency of less common, but potentially more significant terms. To mitigate this issue, the count features are typically re-

scaled into floating-point values that can be effectively employed by a classifier. The commonly used technique to accomplish this is known as "tf-idf" transformation.

### Term Frequency-Inverse Document Frequency (TF-IDF)

The most popular weighting schema is Term Frequency-Inverse Document Frequency (TF-IDF)[101] . It provides each word in a document a weight according to the following two criteria:

- The frequency of its usage in the specified document (TF)
- The rarity of its appearance in the other documents in the corpus (IDF)[102]

Term Frequency-Inverse Document Frequency is the product of two statistics, term frequency and inverse document frequency:

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D) \tag{II. 6}$$

Where t denotes the terms; d denotes each document; D denotes the collection of documents. Term frequency $\mathrm{tf}(t, d)$ is the number of times a term occurs in a document and can be defined as:

$$\mathrm{tf}(d, t) = \frac{f_{td}}{\sum_{t' \in d} f_{t',d}} \tag{II. 7}$$

Where $f_{td}$ is the raw count of a term in a document, i.e., the number of times that term t occurs in document t.
In other words:

$$\mathrm{TF} = \frac{\text{term frequency in document}}{\text{total words in document}} \tag{II. 8}$$

The inverse document frequency is a measure of how much information the term provides, i.e., if it's common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the term (obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient):

$$\mathrm{idf}(t, D) = \log \frac{|\, D \,|}{1 + |\, \{d \in D : t \in d\} \,|} \tag{II. 9}$$

The denominator: $|\, \{d \in D : t \in d\} \,|$ implies the total number of times in which term t appeared in all of a document $d$ (the $d \in D$ restricts the document to be in the current document space). Note that this implies it does not matter if a term appears 1 time or 100 times in a document, it will still be counted as 1, since it simply did appear in the document. As for the plus 1, it is there to avoid zero division.
In other words:

$$\mathrm{idf}(\mathrm{t}, \mathrm{D}) = \log_2 \left( \frac{\text{total documents in corpus}}{\text{documents with term}} \right) \quad \text{(II. 10)}$$

The simpler bag-of-words model and TF-IDF are very simple to understand and implement and offers a lot of flexibility for customization on your specific text data. They have been used with great success on prediction problems like language modeling and documentation classification.

However, it suffers from some shortcomings, such as:

- Vocabulary: The vocabulary requires careful design, most specifically to manage the size, which impacts the sparsity of the document representations.
- Sparsity: Sparse representations are harder to model both for computational reasons (space and time complexity) and for information reasons, where the challenge is for the models to harness so little information in such a large representational space.
- Meaning: Discarding word order ignores the context, and in turn meaning of words in the document (semantics). Context and meaning can offer a lot to the model, that if modeled could tell the difference between the same words differently arranged ("this is interesting" vs "is this interesting"), synonyms ("old bike" vs "used bike"), and much more.

When used with traditional machine learning classifiers, TF-IDF can have high performance on EHR clinical notes[103,104]. As seen on Figure II.2, word and document embedding emerged as the alternative to BoW and TF-IDF offering a representation in a low-dimensional space in which semantically similar tokens are closely positioned as they have similar representations. In contrast to BoW and TF-IDF, word embeddings are represented by a real-valued vector of tens or hundreds of dimensions, having a great performance when representing documents with large vocabulary.

#### *II.2.2.4. Tokenization for deep learning*

Tokenizing a text is splitting it into words or subwords, which then are converted to ids through a look-up table. Converting words or subwords to ids is straightforward and leads to embeddings[2].

##### Word Tokenization

Word tokenization is the most used tokenization algorithm. It splits a piece of text into individual words based on a certain delimiter. Depending upon delimiters, different word-level tokens are formed. Pretrained Word Embeddings such as Word2Vec [105] and GloVe [106] comes under word tokenization. As an example, the **Error! Reference source not found.** shows two types of segmentation, the objective being to divide the raw text into words and find a numerical representation for each word:

[2] An embedding is a numerical representation of a piece of information, for example, text, documents, images, audio, etc. The representation captures the semantic meaning of what is being embedded.

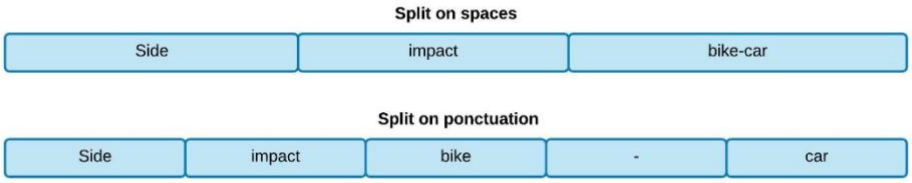



*Figure II.7 Two different ways of splitting a text with word tokenization*

Word tokenization comes along with two major drawbacks:

- Size of vocabulary: The tokenizer employed can result in extensive vocabularies that contain a large number of distinct tokens present in the corpus. Each word is assigned a unique ID, beginning with 0 and extending to the vocabulary size, which is utilized by the model to identify the words. To have complete coverage of a language using a word-based tokenizer, an identifier is required for each word in the language, resulting in an enormous number of tokens. For instance, the English language has more than 500,000 words, necessitating the tracking of that many IDs to establish a mapping from each word to an input ID. Moreover, words such as "dog" and "dogs" are represented differently, and initially, the model will be unaware that they are similar, regarding them as distinct words. Similarly, the model will not initially recognize the similarity between other similar words, such as "run" and "running."
- Out of Vocabulary (OOV) words: When testing, encountering new words that are not present in the vocabulary is known as Out of Vocabulary (OOV) words. These words pose a challenge for the methods used since they lack representation in the vocabulary. To address this, one approach is to establish the vocabulary with the Top K Frequent Words and replace infrequent words in the training data with an unknown token (UNK). This approach enables the model to learn the representation of OOV words utilizing the UNK tokens. Thus, during testing, any word not present in the vocabulary can be mapped to a UNK token. However, this method has certain drawbacks. Firstly, the entire information associated with the OOV word is lost by mapping it to a UNK token. Secondly, each OOV word is assigned the same representation.

### Character Tokenization

Character-based tokenizers split the text into a set of characters, rather than words. This has two primary benefits:

- The vocabulary is much smaller than word-level tokenization
- It handles OOV words by preserving the information of the word.

Character tokens solve the OOV problem but the length of the input and output sentences increases rapidly as we are representing a sentence as a sequence of characters. As a result, it becomes challenging to learn the relationship between the characters to form meaningful words. But there again, there are questions about spacing and punctuation as seen in Figure II.*8*.

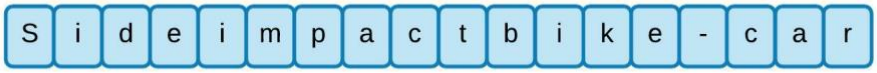



Figure II.8 Character-level tokenization

Although this approach has its drawbacks as well. Rather than using words, the representation is built on characters, which some may argue is less meaningful since each character may not hold significant meaning on its own, unlike words. However, this can vary depending on the language; for instance, Chinese characters convey more information compared to Latin characters.
To achieve a balance between these two approaches, a third technique called subword tokenization can be employed.

### Subword tokenization

Subword Tokenization splits the piece of text into subwords (or n-gram characters). Subword tokenization algorithms rely on the principle that frequently used words should not be split into smaller subwords, but rare words should be decomposed into meaningful subwords. For instance, “walking” might be considered a rare word and could be decomposed into “walk” and “ing”. These are both likely to appear more frequently as standalone subwords, while at the same time the meaning of “walking” is kept by the composite meaning of “walk” and “ing”. This allows to have relatively good coverage with small vocabularies, and close to no unknown tokens.

#### *Byte-Pair Encoding, as used in GPT (Generative Pre-trained Transformer)*

Byte Pair Encoding (BPE)[107] is a simple data compression technique that iteratively replaces the most frequent pair of bytes in a sequence with a single, unused byte. This algorithm was adapted by Sennrich and al.[108] for word segmentation. Instead of merging frequent pairs of bytes, characters or character sequences are merged. When applied to words, the principle is to represent frequent words with fewer symbols and less frequent words with more symbols. BPE is a widely used tokenization method among transformer-based models. BPE addresses the issues of word and character Tokenizers:

- BPE tackles OOV effectively. It segments OOV as subwords and represents the word in terms of these subwords.
- The length of input and output sentences after BPE are shorter compared to character tokenization.

Steps to learn BPE are:

- Split the words in the corpus into characters after appending </w>, each (unicode) character corresponds to a symbol in the final vocabulary.
- Initialize the vocabulary with unique characters in the corpus;
- Compute the frequency of a pair of characters or character sequences in corpus;
- Merge the most frequent pair in corpus;

- Save the best pair to the vocabulary;
- Repeat the frequency of a pair of characters or characters sequences in corpus ;

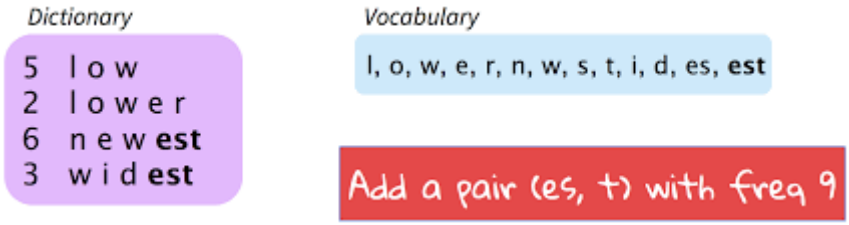



Figure II.9 Byte Pair Encoding with bigrams

BPE tokenizes by merging frequent character n-grams or whole words into a single symbol, removing the need for a shortlist. The final vocabulary size is the initial size plus the number of merge operations, the only hyperparameter. For example, GPT's vocabulary size is 40,478, with 478 base characters and 40,000 merges. However, the model may still encounter unseen characters in the test or supervised training set during pre-training. This can cause embedding vectors for unseen characters to differ significantly from the trained embeddings, even if an ID is created for them. One solution is to create a vocabulary based on a large dataset that includes all necessary characters. Another approach is to use a student-teacher method, where a teacher model is distilled, and a student model is trained using a reduced vocabulary by feeding inputs tokenized with both the student and teacher models[109]. This approach has been used with Bidirectionnal Transformers for Language Understanding[110] (BERT), ELMo[111], and GPT models, and BBPE has also been explored by Wang et al. in 2020[112].

*Byte-level Byte-Pair Encoding (BBPE), as used in GPT2*

When considering all possible base characters, a base vocabulary can become very large, especially if using all Unicode characters. To address this, text can be encoded using UTF-8, which represents each Unicode character with 1 to 4 bytes. However, representing text as a sequence of bytes can result in a longer and more computationally demanding representation compared to a character sequence. To overcome this, byte sequences can be segmented into variable-length n-grams, or byte-level "subwords", using a BPE vocabulary. This extends the UTF-8 byte set with byte n-grams and does not require out-of-vocabulary tokens, enabling language transfer between vocabularies. GPT-2 uses bytes as the base vocabulary to ensure that every base character is included, and its tokenizer can tokenize all text without the need for a special <unk> symbol. Its vocabulary size of 50,257 includes the 256 bytes base tokens, an end-of-text token, and symbols learned with 50,000 merges.

*WordPiece, as used in BERT*

WordPiece is the subword tokenization algorithm used for transformers such as BERT, DistilBERT[113], and Electra[114]. The algorithm was outlined in Japanese and Korean Voice Search [115] and is very similar to BPE.

The WordPiece algorithm is iterative and the summary of the algorithm according to the paper is as follows:

- Initialize the word unit inventory with the base characters (i.e. [CLS]).
- Build a language model on the training data using the word inventory from 1.
- Generate a new word unit by combining two units out of the current word inventory. The word unit inventory will be incremented by 1 after adding this new word unit. The new word unit is chosen from all the possible ones so that it increases the likelihood of the training data the most when added to the model.
- Go to 2 until a pre-defined limit of word units is reached or the likelihood increase falls below a certain threshold.

The only difference between WordPiece and BPE is the way in which symbol pairs are added to the vocabulary. At each iterative step, WordPiece chooses a symbol pair which will result in the largest increase in likelihood upon merging. Maximizing the likelihood of the training data is equivalent to finding the symbol pair whose probability divided by the probability of the first followed by the probability of the second symbol in the pair is greater than any other symbol pair.

### II.2.3 Statistical and Traditional Machine Learning classification algorithms

Some key methods, which are commonly used for text classification are as follows:

- **Decision Trees**: Decision trees are designed with the use of a hierarchical division of the underlying data space with the use of different text features. The hierarchical division of the data space is designed to create class partitions which are more skewed in terms of their class distribution. For a given text instance, we determine the partition that it is most likely to belong to a class and use it for the purposes of classification.

- **Pattern (Rule)-based Classifiers:** In rule-based classifiers we determine the word patterns which are most likely to be related to the different classes. We construct a set of rules, in which the left-hand side corresponds to a word pattern, and the right-hand side corresponds to a class label. These rules are used for the purposes of classification.

- **Support Vector Machines (SVM) Classifiers**: SVM Classifiers attempt to partition the data space with the use of linear or non-linear delineations between the different classes. The key in such classifiers is to determine the optimal boundaries between the different classes and use them for the purposes of classification.

- **Neural Network Classifiers**: Neural networks are used in a wide variety of domains for the purposes of classification. In the context of text data, the main difference for neural network classifiers is to adapt these classifiers with the use of word features. We note that neural network classifiers are related to SVM classifiers; indeed, they both are in the category of discriminative classifiers, which are in contrast with the generative classifiers.

- **Bayesian (Generative) Classifiers**: In Bayesian classifiers (also called generative classifiers), we attempt to build a probabilistic classifier based on modeling the

underlying word features in different classes. The idea is then to classify text based on the posterior probability of the documents belonging to the different classes based on the word presence in the documents.

- **Other Classifiers**: Almost all classifiers can be adapted to the case of text data. Some of the other classifiers include nearest neighbor classifiers, and genetic algorithm-based classifiers. We will discuss some of these different classifiers in some detail and their use for the case of text data.

In the following sections, we describe some of the algorithms for text categorization that have been frequently proposed and used for medical and clinical data mining. First, the general notation is given: Let $\mathbf{d} = \{d_1, \ldots, d_M\}$ be the document vector to be classified and $c_1, \ldots, c_k$ the possible classes. Further assume that we have a training set consisting of $\mathbf{N}$ document vectors $\mathbf{d_1}, \ldots, \mathbf{d_N}$ with true classes $y_1, \ldots, y_N$. $N_j$ is then the number of training documents for which the true class is $i_j$.

### *II.2.3.1. Naïve Bayes*

The naive Bayes classifier[41] is constructed by using the training data to estimate the probability of each class given the document feature values of a new instance. Bayes theorem is used to estimate the probabilities:

$$P(c_j|\mathbf{d}) = \frac{P(c_j)P(\mathbf{d}|c_j)}{P(\mathbf{d})} \qquad (II.11)$$

The denominator in the above equation does not differ between categories and can be left out. Moreover, the naive part of such a model is the assumption of word independence, i.e we assume that the features are conditionally independent given the class variable. This simplifies the computations yielding:

$$P(c_j|d) = P(c_j)\prod_{i=1}^{M} P(d_j|c_j) \qquad (II.12)$$

An estimate $\widehat{P}(c_j)$ for $P(c_j)$ can be calculated from the fraction of training documents that is assigned to class $c_j$ :

$$P(C = c_j) = \frac{N_j}{N} \qquad (II.13)$$

Moreover, an estimate $\widehat{P}(d_j|c_j)$for $P(d_j|c_j)$ is given by:

$$\widehat{P}(d_j|c_j) = \frac{1 + N_{ij}}{M + \sum_{k=1}^{M} N_{kj}} \quad (II.14)$$

where $N_{ij}$ is the number of times word i occurred within documents from classe $c_j$ in the training set.
Even though the assumption of conditional independence is generally not true for word appearance in documents, the Naive Bayes classifier is surprisingly effective[116,117].

#### *II.2.3.2. K-nearest neighbor (KNN)*

To classify an unknown document vector $\mathbf{d}$, the k-nearest neighbor (kNN) algorithm[118] ranks the document's neighbors among the training document vectors and use the class labels of the $\mathbf{k}$ most similar neighbors to predict the class of the input document. The classes of these neighbors to predict the class of the input document. The classes of these neighbors are weighted using the similarity of each neighbor to $\mathbf{d}$, where similarity may be measured by Euclidean distance or the cosine between the two document vectors for example.
KNN is a lazy learning instance-based method that does not have an off-line training phase. The main computation is the on-line scoring of training documents given a test document to find the $\mathbf{k}$ nearest neighbors.

#### *II.2.3.3. Decision Trees*

In this approach, the document vector $\mathbf{d}$ is matched against a decision tree to determine whether the document is relevant for a given category. The decision tree is constructed from the training samples, and one of the most popular approaches is the CART (Classification And Regression Trees)[119] algorithm that is described below.

##### CART

CART builds a binary decision tree by splitting the set of training vectors at each node according to a function of one single vector element. The first task is therefore to decide which of the vector elements makes the best splitter, i.e the one that partitions the set in as homogeneous subsets as possible. This means that the best splitter is the one that decreased the diversity of the set of training samples by the greatest amount, i.e one wants to maximize:

$$\text{diversity(before split)} - [\text{diversity(left child)} + \text{diversity(right child)}]$$

One of the commonly used diversity measures is entropy:

$$\sum_{j=1}^{k} p(c_j|t)\log p(c_j|t) \quad (II.15)$$

Where $p(c_j|t)$ is the probability of a training sample being in class $c_j$ given that it falls into node t. This probability can be estimated by:

$$p(c_j|t) = \frac{N_j(t)}{N(T)} \quad (II.16)$$

where $N_j(t)$ and $N(t)$are the number of samples of class $c_j$ and the total number of samples at node t respectively.

To choose the best splitter at a node in the tree, each component of the document vector is considered in turn. A binary search is performed to determine the best split value for the component, using the decrease in diversity as the measure of goodness. Having found the best split value, one compares the decrease in diversity to that provided by the current best splitter. The component that matches to the largest decrease in diversity is chosen as the splitter for the node.

This procedure is repeated until no sets can be partitioned any further. The nodes at the bottom of the tree are denoted leaf nodes, and at the end of the tree-growing process, every sample of the training set has been assigned to some leaf of the full decision tree.

Each leaf can now be assigned a class. The error rate of a leaf measures the probability of samples reaching this leaf being misclassified. The error rate, $E(T)$, of the whole tree is the weighted sum of the error rates of all the leaves.

#### Other Algorithms

Two other well-known decision tree algorithms are C4.5[120] and CHAID (CHi-squared Automatic Interaction Detector)[121].

C4.5 differs from CART in that it produces trees with varying numbers of branches per node.

CHAID differs from CART and C4.5 in that rather than first overfitting the data, then pruning, CHAID attempts to stop growing the tree before overfitting occurs. CHAID is restricted to categorical variables.

### *II.2.3.4. Support Vector Machines (SVM)*

Support Vector Machines (SVMs) have shown to yield good generalization performance on a wide variety of classification problems. The SVM integrates the dimension reduction and classification. In a binary classification task, the SVM classifies a vector $\mathbf{d}$ to either -1 or 1 using:

$$s = \mathbf{w}^T \emptyset(\mathbf{d}) + b = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{d}, \mathbf{d}_i) + b \quad (II.17)$$

and

$$y = \begin{cases} 1 \text{ if } s > s_0 \\ -1 \text{ otherwise} \end{cases} \quad (II.18)$$

Here $\{\mathbf{d}_i\}_{i=1}^{N}$ is the set of training vectors as before and $\{y_i\}_{i=1}^{N}$ are the corresponding classes $(y_i \in -1,1)$ is denoted a kernel and is often chosen as a polynom of degree $d$, i.e.

$$K(d, d_i) = (d^T d_i + 1)^d \qquad (II.19)$$

The training of the SVM consists of determining the $\mathbf{w}$ that maximizes the distance between the training samples from each pair of class.

### *II.2.3.5. Voted Classification*

Many researchers have investigated the technique of combining the predictions of multiple classifiers to produce a single classifier. This process is often denoted voting. Voting algorithms takes a classifier and a training set as input and trains the classifier multiple times on different versions of a training set. The generated classifiers are then combined to create a final classifier that is used to classify the test set.
Voting algorithm can be divided into two types: bagging and boosting algorithms. The main difference between the two types is the way the different versions of the set are created. We give, in the following, a closer description of the two types of algorithms.

#### Bagging

Bagging[122] takes as input a classification algorithm $f(.)$ and a training set $T$ and returns a set of classifiers $f^*(.) = f_1(.), \ldots\ldots, f_R(.)$. Here $f_r(.)$ is a classifier that is learned from a bootstrap sample $T_r$ of the training set. The bootstrap sample is formed by uniform probability random selection from $T$ with replacement $N$ times, where $N$ is the size of the training set. This will create a training set with the same number of samples as the original, but some cases may be represented more than once, while others may not be represented at all. The expected frequency with which the cases from $T$ are represented in a single bootstrap sample $T_r$ is described in the discrete Poisson distribution.
To classify a new sample $\mathbf{d}$, each classifier $f_r(.)$ from \$ $f^*(.)$ is applied to $\mathbf{d}$ resulting in labels $f_1(\mathbf{d}), f_2(\mathbf{d}), \ldots\ldots, f_R(\mathbf{d})$ . The result of the voting classifier is the class that obtains the most votes from the single classifiers when applied to $\mathbf{d}$:

$$f^*(\mathbf{d}) = \text{argmax}_y \sum_{r:f_r(\mathbf{d})=y} 1 \qquad (II.20)$$

#### Boosting

Boosting[123] encompasses a family of methods. Like bagging, these methods choose a training set of size $N$ for classifier f_r by randomly selecting with replacement examples from the original training set. Unlike bagging, however, the probability of selecting a sample is not the same for all samples of the training set. It depends instead on how often that sample was misclassified by the previous $k-1$ classifier. Thus, boosting attempts to produce new

classifiers that are better able to correctly classify examples from which the performance of the current classifiers is poor.
Different forms of boosting generate the probabilities for selecting samples in different ways. We will describe two approaches here: the AdaBoost, AdaBoost.MH and gradient boosting.

### *AdaBoost*

Let the probability of selecting the sample $\mathbf{d}_i$ for training set $T_r$ be $p_{ir}$. Initially, all the probabilities are equal, i.e. $p_{ir} = 1/N$ for all samples $\mathbf{d}_i$. To determine the $p_{ir}$'s for classifier $f_{r+1}(.)$ AdaBoost first computes the sum, $\varepsilon_r$, of the probabilities corresponding to the samples that were misclassified using classifier f_r(.):

$$\varepsilon_r = \sum_{i:f_r(d_i)\neq y_i} p_{ir} \qquad (II.21)$$

Finally, the probabilities are re-normalized so that they again sump up to 1.
After this procedure has been repeated for R iterations, R classifiers $f_1(.), \ldots., f_r(.)$ and R values $\alpha_1, \ldots, \alpha_R$ remain. To classify a new sample $\mathbf{d}$, each classifier $f_r(.)$ from $f^*(.)$ is applied to $\mathbf{d}$ resulting in labels $f_1(\mathbf{d}), f_2(\mathbf{d}), \ldots \ldots, f_R(\mathbf{d})$. Unlike bagging, one does not assign equal importance to each of the classification results, but instead weight the results using the $\alpha_r$ values that were previously used to update the probabilities $p_{ir}$. This means that the final class $\mathbf{d}$ is given by:

$$f^*(\mathbf{d}) = \text{argmax}_y \sum_{k:f_r(\mathbf{d})=y} \alpha_r \qquad (II.22)$$

A main mis advantage with AdaBoost is that is not very good at solving multi-class problems. In addition, it doesn't handle cases where a document may belong to more than one class. An extension of AdaBoost called AdaBoost.MH[124] can effectively handle multi-class and multi-label problems.

### *AdaBoost.MH*

Let the weight of sample $\mathbf{d}_i$ and label $c_k$ in iteration r be $p_{ikr}$ . Initially, all weights are equal, i.e $p_{ikr} = \frac{1}{N}$ for all samples $\mathbf{d}_i$ and all labels $c_k$. For each round, the AdaBoost.MH algorithm estimates K classifiers $f_r(d, k)$. The sign of $f_r(d_i, k)$reflects whether the label $c_k$ is or is not assigned to the training sampe $\mathbf{d}_i$, while the magnitude of $f_r(d_i, k)$ is interpreted as a measure of the confidence in the prediction. The weights are updated using the following formula:

$$p_{ik(r+1)} = p_{ikr}\exp\left(-y_{ik}f_r(d_i, k)\right) \qquad (II.23)$$

Here $y_{ik}$ is 1 if label$c_k$ is among the possible true labels of sample $\mathbf{d}_i$ and -1 otherwise. After updating the weights, they are re-normalized so that $\sum_i \sum_{kp\_\{ik(r+1)\}} = 1$.

After this procedure has been repeated for R iterations, one has RxK classifiers $f_r(\mathbf{d}, k)$. To classify a new sample $\mathbf{d}$, each classifier is applied to $\mathbf{d}$ and the final class of $\mathbf{d}$ is given by:

$$f^*(\mathbf{d}, k) = \sum_{r=1}^{R} f_r(\mathbf{d}, k) \qquad (II.24)$$

*GradientBoosting*

AdaBoost and related algorithms were recast in a statistical framework first by Breiman[122] calling them ARCing (Adaptive Reweighting and Combining) algorithms. Each step in an arcing algorithm consists of a weighted minimization followed by a recomputation of the classifiers and weighted input. While the AdaBoost model identifies the shortcomings by using high weight data points, gradient boosting performs the same by using gradients in the loss function.

#### *II.2.3.6. Traditional machine learning classifiers for EHR clinical notes*

Since the beginning of computerized EHR, traditional machine learning has been widely used for extracting information from clinical notes. Many algorithms have been tested on various classifying tasks and some trends have been drawn regarding how to perform on clinical notes. As mentioned in section II.1.2.2, SVM has been found by Wang et al.[125] to be the most frequently used algorithm. When considering NLP techniques for chronic disease, E. H. Houssein et al. showed in their systematic review that SVM and Naïve Bayes were widely used[126]. Barrett et al. [127] found similar results while comparing SVM, Naïve Bayes and Maximum Entropy for adverse drug reaction detection with a classification task. Similar results were found by Sarker et al.[127] for the same task. Roberts et al.[128] proposed an approach to use SVM with various features to extract anatomic sites of appendicitis-related findings.
As clinical notes have high-dimensional feature spaces and sparse instance vectors, these issues were found to be well addressed by SVMs[41]. They have been recognized for their generalizability and are largely used for phenotyping[54]. On the other hand, Metzger et al., while classifying French ED clinical notes for suicide with a multi-class classification, compared several traditional machine learning models and neural network and found random forest to be the most accurate model with a 95.3 F-measure, while SVM reached 90.4[62].

### II.2.4 Deep Learning versus Statistical and Traditional Machine Learning

Deep learning is a class of machine learning algorithms based on artificial neural networks with representation learning[129] as seen on Figure II.*10*. Machine learning is defined as the study of computer algorithms that improve automatically through experience which is seen as a subset of artificial intelligence[130]. For a long time, the majority of methods used to study NLP problems employed shallow machine learning models and time-consuming, hand-crafted features. This led to problems such as the curse of dimensionality since linguistic information was represented with sparse representations (high-dimensional features). However, with the recent popularity and success of word embeddings (low dimensional, distributed

representations), neural-based models have achieved superior results on various language-related tasks as compared to traditional machine learning models.

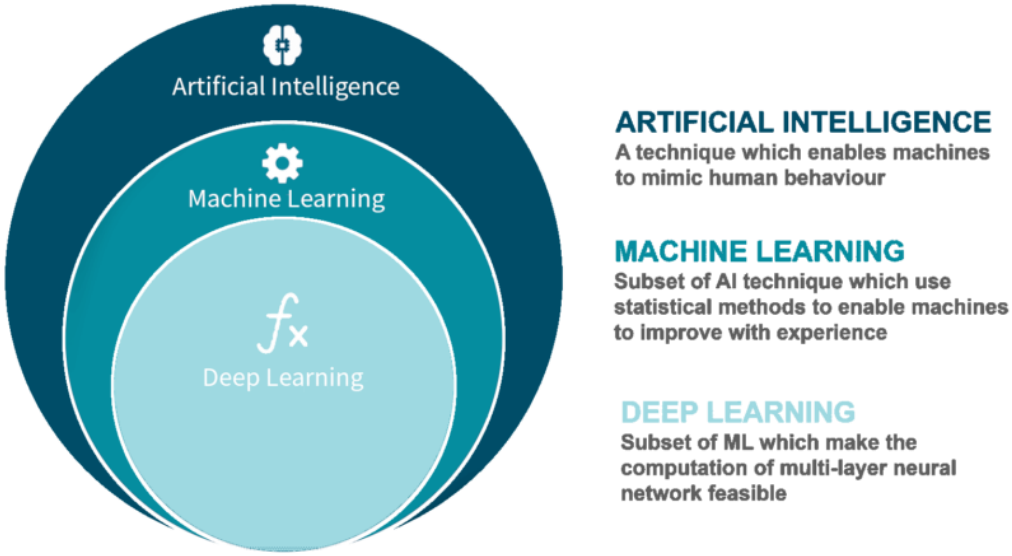



Figure II.10 Artificial Intelligence, Machine Learning and Deep Learning connections

Deep learning uses multiple layers to progressively extract higher level features from the raw input. For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces[131].

### II.2.5 Artificial Neural Networks

Artificial neural networks form the core of deep learning applications. Neural networks are arrangements of multiple nodes or neurons, arranged in multiple layers as seen on Figure II.11.

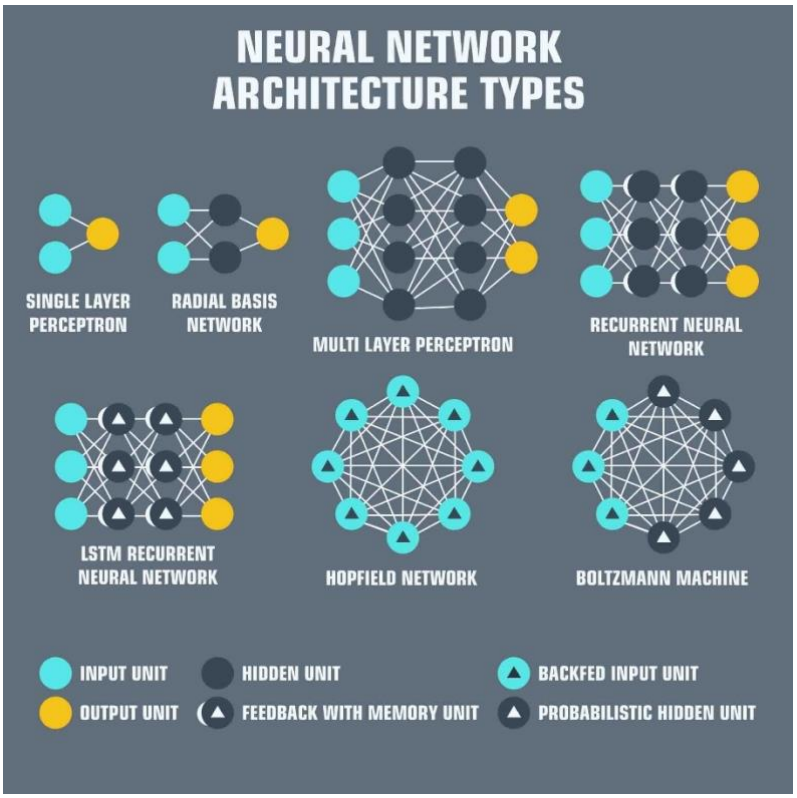



*Figure II.11 Types of Artificial Neural Networks*

The information enters the neural network through the input layer, which is the primary outermost layer. The final layer through which information passes is the output layer. The input and output layers may or may not have additional layers between them. The layers, if any, present between the input and output layers are called hidden layers. An artificial neural network is considered to be a "deep" neural network if it has multiple hidden layers. Generally, every neuron in any layer is interconnected to all neurons in its adjacent layers.

Every layer of the neural network breaks down the input into a simpler form to interpret and classify the content. For instance, consider a simple neural network that is used to identify the pictures of cats. The different layers of the neural network perform different functions and analyze different elements of input images. For example, the first later could simply scan for contours in the images. The next layer can identify different colors. Similarly, the subsequent layers can make increasingly detailed analyses to identify more subtle features, ultimately allowing the neural network to identify the images of cats distinctly. A high number of layers means that there can be a higher number of pathways for information to travel through the network, potentially allowing the network to perform highly complex tasks.

Until recently, the 3 most commonly used types of neural networks in AI were:

- Feed Forward Neural Networks: used to perform basic pattern and image recognition
- Convolutional Neural Networks (CNNs): used in object recognition and video analysis
- Recurrent Neural Networks (RNNs): used in Natural Language Processing and speech recognition

In 2017, the Transformer architecture revolutionized NLP, and the recent popularity of ChatGPT has not only given visibility to this type of model but has opened the field of possibilities in a non-finite number of areas.

### *II.2.5.1. Convolutional Neural Network (CNN)*

Convolutional Neural Network (CNNs) are deep learning models that are most used for computer vision applications (such as classifying images), and in some cases used for natural language processing tasks (such as text classification). A CNN is made up of an input layer, hidden layers, and an output layer. The hidden layers in a CNN include one or more layers that perform convolutions. A convolution layer computes a dot product of the convolution kernel with the layer's input matrix, usually using a ReLU activation function. The convolution kernel slides along the input matrix for the layer, generating a feature map which contributes to the input of the next layer. Other layers such as pooling layers, fully connected layers, and normalization layers may follow.

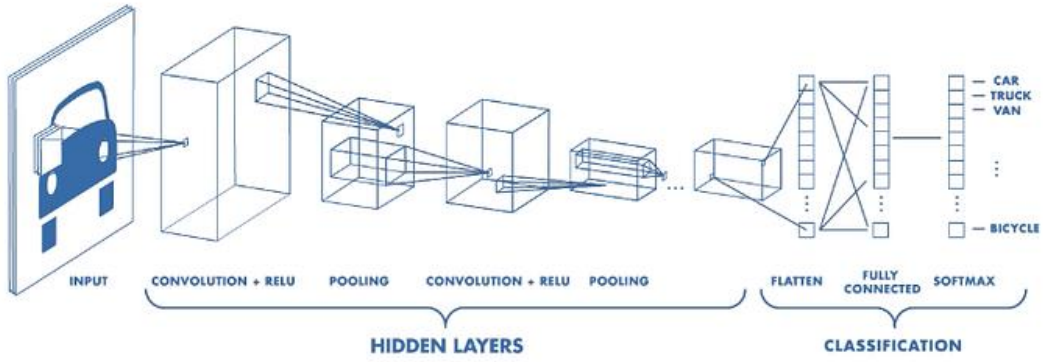



*Figure II.12 Convolutional Neural Network*[132]

#### Convolution Layer

The convolution layer is the core building block of the CNN. It carries the main portion of the network's computational load. This layer performs a dot product between two matrices, where one matrix is the set of learnable parameters otherwise known as a kernel, and the other matrix is the restricted portion of the receptive field.

In the case of NLP tasks, i.e., when applied to text instead of images, we have a 1-dimensional array representing the text. Here the architecture of the CNN is changed to 1D convolutional-and-pooling operations. One of the most typically tasks in NLP where CNN are used is sentence classification, that is, classifying a sentence into a set of pre-determined categories by considering $n$-grams, i.e. it's words or sequence of words, or also characters or sequence of characters.

Given a sequence of words $w_{1:n} = w_1, \dots, w_n$ , where each is associated with an embedding vector of dimension $d$. A 1D convolution of width-$k$ is the result of moving a sliding-window of size $k$ over the sentence and applying the same convolution filter or kernel to each window in the sequence, i.e., a dot-product between the concatenation of the embedding vectors in a given window and a weight vector $u$, which is then often followed by a non-linear activation function $g$.

Considering a window of words, $w_i, \dots, w_{i+k}$ the concatenated vector of the ith window is then:

$$x_i = [w_i, w_{i+1}, \dots, w_{i+k}] \in \mathbb{R}^{k \times d} \quad (II.25)$$

The convolution filter is applied to each window, resulting in scalar values $r_i$ , each for the ith window:

$$r_i = g(x_i \cdot u) \in \mathbb{R} \quad (II.26)$$

In practice one typically applies more filters, $u_1, \dots, u_l$ , which can then be represented as a vector multiplied by a matrix $U$ and with an addition of a bias term $b$ :

$$r_i = g(x_i \cdot U + b) \quad (II.27)$$

With

$$r_i \in R^l, \quad x_i \in R^{k \times d}, \quad U \in R^{k \cdot d \times l} \text{ and } b \in R^l \quad (II.28)$$

### Pooling

The pooling operation is used to combine the vectors resulting from different convolution windows into a single $l$-dimensional vector. This is done again by taking the max or the average value observed in resulting vector from the convolutions. Ideally this vector will capture the most relevant features of the sentence/document. This vector is then fed further down in the network, hence, the idea that CNN itself is just a feature extractor, most probably to a full connected layer to perform prediction.

### CNN for EHR clinical notes classification

Deep models such as CNNs have attracted attention and achieved very competitive results in classification tasks. One of the first attempts was by Hughes et al[133], who applied a CNN to classify clinical text at the sentence level. The model structure had four convolutional layers after the sentence embedding input, and at the end a fully connected layer was applied to predict the sentence labels. They compared their method with a variety of traditional machine learning methods and different sentence embeddings, including logistic regression, doc2vec embeddings, and bag-of-words features. Results from Hughes et al. and others have demonstrated that deep models, including word embeddings and CNN models, are competitive with, and can even outperform, TF-IDF and topic modeling features[133,134].

### *II.2.5.2. Recurrent Neural Network (RNN)*

#### Overview of RNN

RNNs have been applied to various NLP tasks, including machine translation, image captioning, and language modeling, among others. Compared to CNN models, RNN models can be equally effective or even better at specific natural language tasks but not necessarily superior, as they model different aspects of the data, depending on the semantics required by the task at hand. However, simple RNNs are challenged by the vanishing gradient problem, which makes learning and tuning the parameters in the earlier layers difficult.
Recurrent Neural Networks (RNNs) utilize loops to perform recurrent operations, which make them more intricate than feed-forward networks and capable of tackling complex tasks such as language generation and text prediction. Unlike feed-forward networks, RNNs allow connections to go back to neurons in the same layer, which widens the scope of operations. RNNs perform recursive computations for every instance of an input sequence based on the previous computed results. Typically, these sequences are represented by a fixed-size vector of tokens that are fed sequentially to a recurrent unit. A simple RNN framework is illustrated in the Figure II.*13* below.

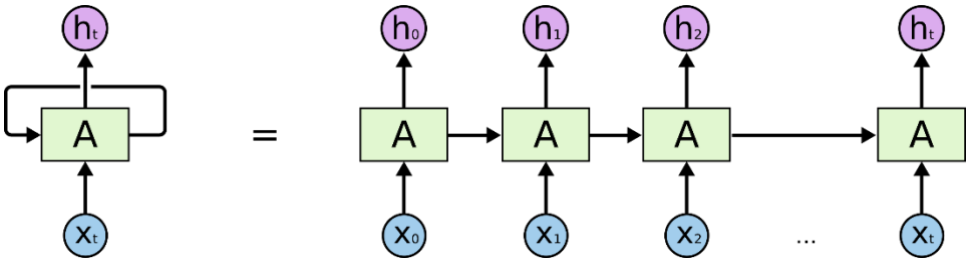



Figure II.13 Unrolled RNN[135]

RNNs are designed to take two inputs at each time step: an input $x_t$ and a hidden state $h_t$. The second input vector and the first hidden state are used to generate the output of that time step. RNNs are particularly useful for modeling context dependencies in inputs of arbitrary length and creating an appropriate composition of the input by memorizing the results of previous computations and using that information in the current computation.

## Types of RNN

Five main types of RNN are depicted on Figure II.*14*:

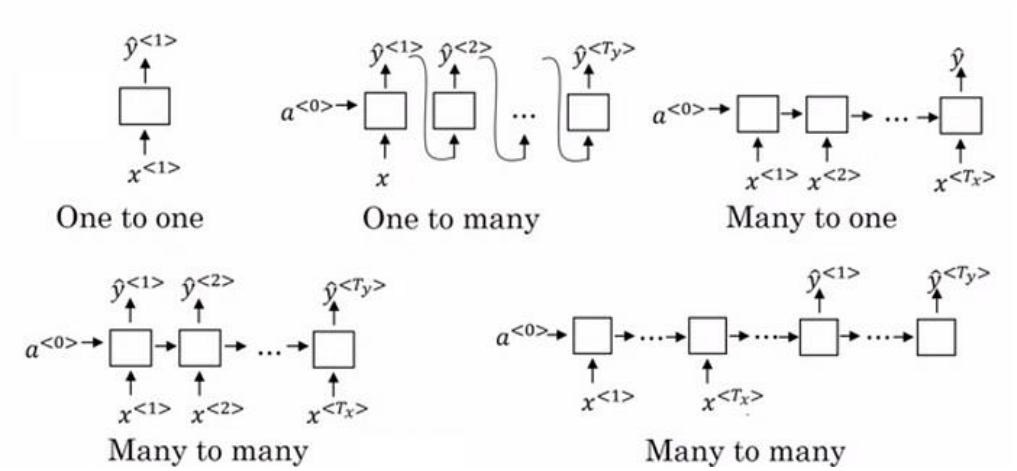



Figure II.14 Types of RNN

- One to One RNN is the most basic and traditional type of Neural network giving a single output for a single input, as can be seen in the above image. It is also known as Vanilla Neural Network. It is used to solve regular machine learning problems.
- One to Many is a kind of RNN architecture is applied in situations that give multiple output for a single input. A basic example of its application would be Music generation. In Music generation models, RNN models are used to generate a music piece (multiple output) from a single musical note (single input).
- Many to One RNN architecture is usually seen for sentiment analysis model as a common example. As the name suggests, this kind of model is used when multiple inputs are required to give a single output.
- Many-to-Many RNN architecture takes multiple input and gives multiple output, but Many-to-Many models can be two kinds as represented above:
    1. When input and output layers have the same size. This can be also understood as every input having a output, and a common application can be found in NER.
    2. Many-to-Many architecture can also be represented in models where input and output layers are of different size, and the most common application of this kind of RNN architecture is seen in Machine Translation. For example, “He fell” in English is translated to 3 words in French, “Il est tombé”. Thus, machine translation models can return words more or less than the input string because of a non-equal Many-to-Many RNN architecture works in the background.

## RNN Drawbacks

### *Gradient definition*

The gradient is a is a partial derivative with respect to its inputs. In other words, it describes the slope or rate of change of a function with respect to its input parameters. Specifically, the gradient of a function is a vector that points in the direction of steepest ascent of the function

at a particular point in its domain. Higher the gradient, steeper the slope and the faster a model can learn. If the slope is almost zero, the model stops to learn. A gradient simply measures the change in all weights regarding the change in error.
In the context of deep learning, the function of interest is the loss function, which measures the discrepancy between the predicted output of a neural network and the true output. The gradient of the loss function with respect to the network's parameters (i.e., the weights and biases of its neurons) tells us how much each parameter should be adjusted to reduce the loss.
To compute the gradient of the loss function, backpropagation is used, which involves recursively applying the chain rule of calculus to propagate the derivatives of the loss function through the layers of the network. The resulting gradient vector can then be used to update the network's parameters using an optimization algorithm such as stochastic gradient descent (SGD).

#### *Gradient issues in RNN*

During the training of an RNN algorithm, it is possible for the gradient to become either too small or too large. As a result, the training process of the algorithm can become challenging. This can lead to various issues such as poor performance, low accuracy, and an extended training period.

##### *Exploding Gradient*

If we assign significant importance to the weights in a neural network, it can result in the problem of an exploding gradient. This occurs when the gradient values become excessively large, causing the slope to grow exponentially. To address this issue, several techniques can be used, including identity initialization, truncated back-propagation, and gradient clipping.

##### *Vanishing Gradient*

The problem of vanishing gradient[136] occurs when the gradient values become too small during model training, causing the learning process to either slow down significantly or come to a halt. To overcome this issue, several methods can be employed, including weight initialization, selecting appropriate activation functions, and incorporating gating mechanisms. Gating mechanisms have been developed to alleviate some limitations of the basic RNN, resulting in two prevailing RNN types: long short-term memory (LSTM)[137] and gated recurrent unit (GRU)[138].

### RNNs for EHR clinical notes classification

As highlighted by Wang et al.[125], RNNs (LSTM comprised) are the most frequently algorithms used for information extraction from EHR clinical notes. As an example, Futoma et al. [139] developed an RNN classifier to detect sepsis in EHRs. They framed the problem as a multivariate time series classification problem. By using a multitask Gaussian process and feeding into an RNN, they were able to achieve significant improvements over baselines and clinical benchmarks with significantly higher precision. Specifically, compared with vanilla RNN, the model improved the precision by 0.1; compared with non-deep learning methods, the model improved the precision by 0.3-0.4.

In 1997, Hochreiter and Schmidhuber introduced LSTM[137] to address the long-term dependency problem.
LSTM rely on 2 states, the hidden state and the cell state and 3 gates:

- Forget Gate (ability to forget information when it is no longer useful)
- Input Gate (ability to consider new relevant information)
- Output Gate (determines the state of the cell at time t, given the forget gate and input gate)

Let's denote:
- $x_t \in \mathbb{R}^d$ : input vector to the LSTM unit
- $f_t \in (0,1)^h$ : forget gate's activation vector
- $i_t \in (0,1)^h$: input/update gate's activation vector
- $o_t \in (0,1)^h$ : output gate's activation vector
- $h_t \in (-1,1)^h$: hidden state vector also known as output vector of the LSTM unit
- $\tilde{c}_t \in (-1,1)^h$: cell input activation vector
- $c_t \in \mathbb{R}^d$: cell state vector
- $W \in \mathbb{R}^{h \times d}$, $U \in \mathbb{R}^{h \times h}$ and $b \in \mathbb{R}^h$: weight matrices and bias vector parameters which need to be learned during training where the superscripts $d$ and $h$ refer to the number of input features and number of hidden units, respectively.

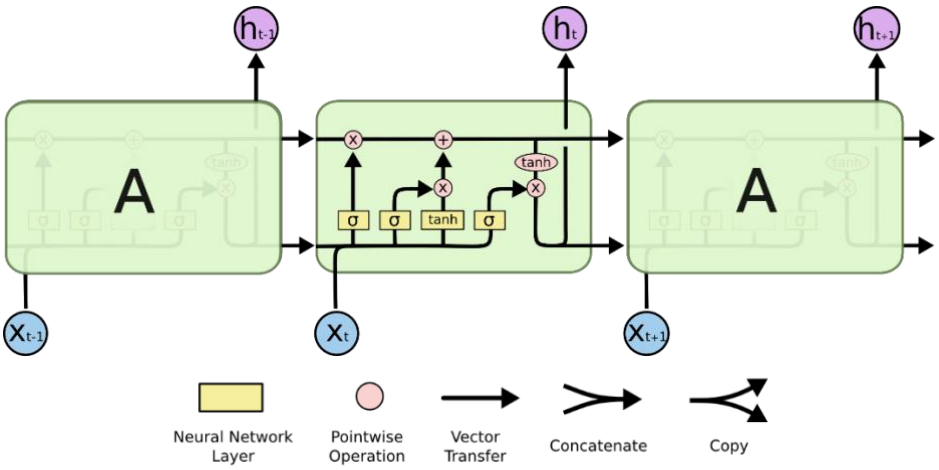



Figure II.15 LSTM units with the 4 interacting layers[135]

In an LSTM neural network, a sigmoid layer as seen on Figure II.*16* is used to control the flow of information through the cell state Figure II.*17*. The sigmoid function is a mathematical function that maps any input value to a value between 0 and 1.

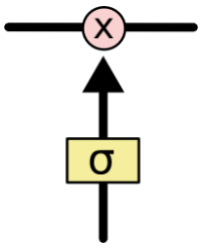



Figure II.16 Sigmoid layer of an LSTM cell[135]

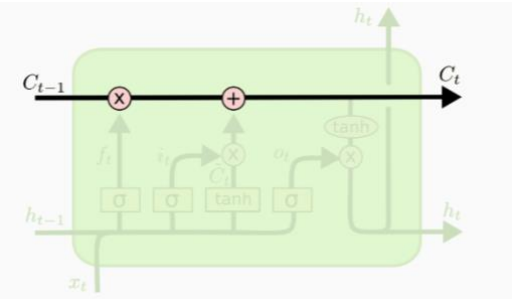
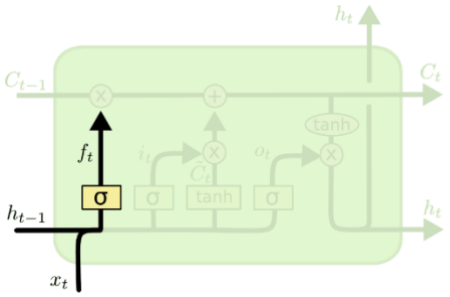


Figure II.17 Cell state of an LSTM[135]

Specifically, there are three sigmoid layers used in an LSTM network:

1. The forget gate (Figure II.*18*): This gate determines how much of the previous cell state $c_{t-1}$ should be forgotten. The forget gate takes as input the current input $x_t$ and the previous hidden state $h_{t-1}$, and outputs $f_t$ a value between 0 and 1 for each element in the cell state $c_t$.



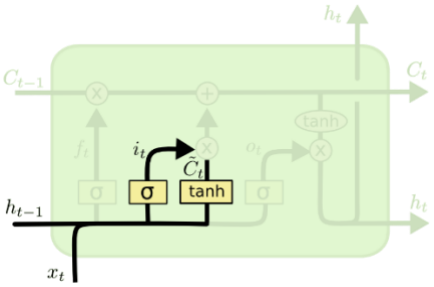
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Figure II.18 Forget gate layer of LSTM[135]

2. The input gate (Figure II.*19*): The input gate is a sigmoid layer that takes as input the current input $x_t$ and the previous hidden state $h_{t-1}$. The output $i_t$ of the input gate determines how much new information should be added to the cell state $c_t$. Specifically, the input gate output $i_t$ is multiplied elementwise with the output of the tanh layer $\tilde{c}_t$.

   The tanh layer is another layer in an LSTM that takes as input the current input $x_t$ and the previous hidden state $h_{t-1}$. The output of the tanh layer $\tilde{c}_t$ is a value between -1 and 1 for each element in the cell state $c_t$. This product output of the input gate $i_t$ is multiplied elementwise with the output of the tanh layer $\tilde{c}_t$ determines how much new information should be added to the cell state. If the output of the input gate is close to 0, then very little new information will be added to the cell state. If the output of the input gate $i_t$ is close to 1, then a lot of new information will be added to the cell state $c_t$.

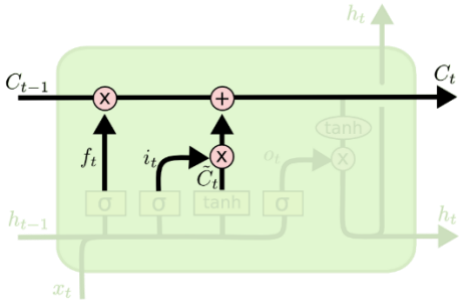
   The purpose of the tanh layer is to normalize the new input before it is added to the cell state. This is important because the cell state can grow very large or small over time, which can make it difficult to train the network. By using the tanh function, the new input is scaled to a range between -1 and 1, which helps to prevent the cell state from growing too large or too small.




$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Figure II.19 Input gate of LSTM[135]

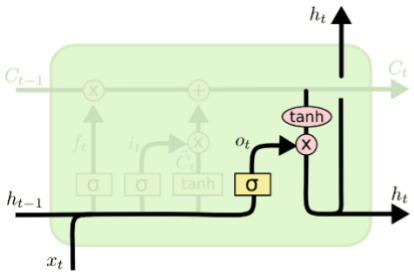
3. The output gate is the final layer that produces the output of the network.
Before the activation of the output gate, the old cell state, $c_{t-1}$ is updated into the new cell state $c_t$ as seen on Figure II.*20*. The previous steps already decided what to do, we just need to actually do it. The old state $c_{t-1}$ is multiplied by $f_t$, forgetting the unnecessary information, then $\tilde{c}_t$ is added resulting in the new candidate values scaled by how much each value state was updated.




$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Figure II.20 Output gate of LSTM[135]

Finally, the output layer takes as input the previous hidden state $h_{t-1}$ and applies a linear transformation to this input. The output of the output layer is then passed through an activation function, which is often a softmax function in the case of sequence classification or language modeling, but here is a tanh function. The softmax or tanh function normalizes the output into a probability distribution over a set of output classes. In some cases, the output layer may include additional layers or regularization techniques, such as dropout or batch normalization, to improve the performance of the network.




$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Figure II.21 Output gate of LSTM[135]

GRU

GRU (Gated Recurrent Unit) is a type of recurrent neural network that is similar to LSTM (Long Short-Term Memory) but has fewer parameters and is faster to train. Like LSTM, GRU is designed to model sequential data by processing input sequences of variable length. GRU works by using gating mechanisms to control the flow of information through the network. The basic GRU unit consists of two gates: an update gate and a reset gate.

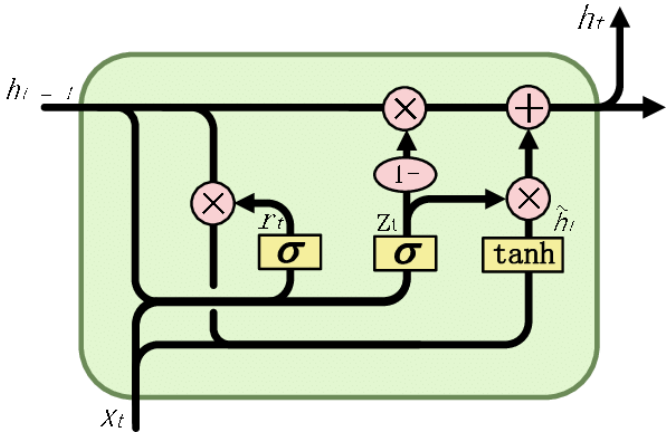



Figure II.22 GRU structure $z_t$ and $r_t$ represent the update gate and reset gate respectively[140]

The update gate $\mathrm{z_t}$ determines how much of the previous hidden state $\mathrm{h_{t-1}}$ should be retained and how much of the new input should be used to update the hidden state $\mathrm{h_t}$. The reset gate $\mathrm{r_t}$ determines how much of the previous hidden state $\mathrm{h_{t-1}}$ should be forgotten. At each time step, the input $\mathrm{x_t}$ is concatenated with the previous hidden state $\mathrm{h_{t-1}}$, and the resulting vector is passed through a linear layer. The output of this layer is then split into two vectors, one of which is used to compute the reset gate, and the other is used to compute the update gate.
The reset gate $\mathrm{r_t}$ is computed using a sigmoid activation function, which maps the input to a value between 0 and 1. The update gate is also computed using a sigmoid function, which determines how much of the new input should be used to update the hidden state.
Once the reset and update gates have been computed, the previous hidden state is multiplied by the reset gate. This multiplication essentially determines how much of the previous hidden state should be forgotten. The new input is then passed through a tanh activation function, which scales the input to a value between -1 and 1.
Finally, the scaled input and the previous hidden state multiplied by the reset gate are combined to compute the new hidden state. The update gate determines how much of the new input should be used to update the hidden state, and the remaining proportion of the previous hidden state is used to maintain the long-term memory of the network.
GRU is a simpler and faster alternative to LSTM that is often used for modeling sequential data, such as natural language processing, speech recognition, and video analysis.

### *II.2.5.3. Attention Mechanism*

Attention, as introduced is a mechanism in deep learning that allows a model to focus on certain parts of the input while processing it. The attention mechanism works by assigning a weight to each input element based on its relevance to the current output. These weights are then used to compute a weighted sum of the input elements, which is used as a contextual representation for the current output.

The attention mechanism typically consists of three components:

1. Query: The current output of the model, which is used to compute the weights for each input element.
2. Key: The set of input elements, which are used to compute the weights for each element.
3. Value: The set of input elements, which are weighted and combined to produce the contextual representation for the current output.

To compute the weights for each input element, the attention mechanism typically uses a scoring function, which takes the query and the key as inputs and produces a scalar score. The score is then transformed into a weight using a softmax function, which ensures that the weights sum to one.

There are several types of scoring functions that can be used for the attention mechanism, including dot product, additive, and multiplicative. Each has its own advantages and disadvantages depending on the task at hand.

Once the weights have been computed, they are used to compute a weighted sum of the values, which produces the contextual representation for the current output. This contextual representation can then be used as input to the next layer of the model, or as the final output of the model.

#### *II.2.5.4. Transformers*

In recent years, the rapid development of Large language Models (LLMs) based on Transformer architecture has been revolutionizing the field of NLP[141]. We will explain in detail the vanilla Transformer architecture, self-attention, unsupervised pre-training. Then, we will go in depth into BERT and GPT models.

##### The transformer architecture

Introduced in 2017 by Google and proposed in the article “Attention is All You Need” by Vaswani et al[66], transformers have an architecture that allows the implementation of a mechanism for processing the sequence of tokens that form a sentence in a self-attentive manner, that is, relating each of these tokens to each of the others in the sentence. They have the particularity of being able to be pretrained on a corpus of text, which can be very large because it does not require a coding stage. This phase leads to a generative model that is capable, for example, of constructing artificial text by iteration.

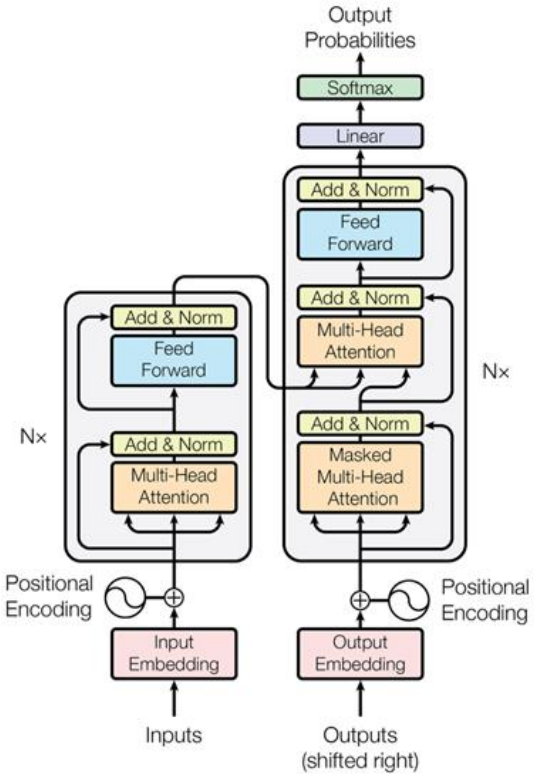



Figure II.23 The transformer architecture proposed by Vaswani et al.[66]

The major components of a Transformer are the encoder and decoder. The encoding component is a stack of encoders (6 in the original architecture). The decoding component is a stack of decoders of the same number as seen on Figure II.*24*.

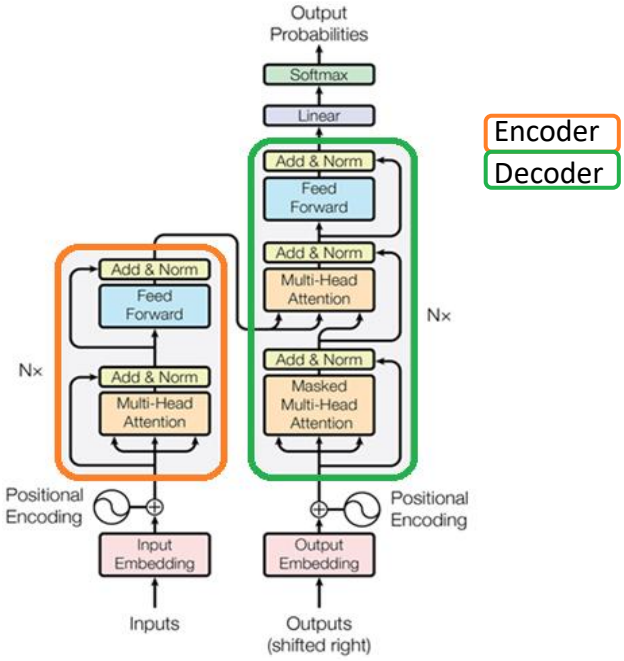



Figure II.24 Encoder/decoder components of the Transformer[66].

*Figure II.25 Encoder component of the Transformer*[142]

The encoders are all identical in structure. The encoder's inputs first flow through a self-attention layer – a layer that helps the encoder look at other words in the input sentence as it encodes a specific word). The outputs of the self-attention layer are fed to a feed-forward neural network. The exact same feed-forward neural network (FFNN) is independently applied to each position.

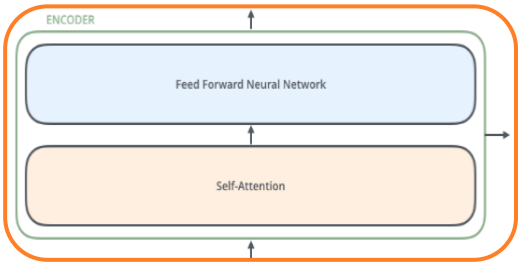
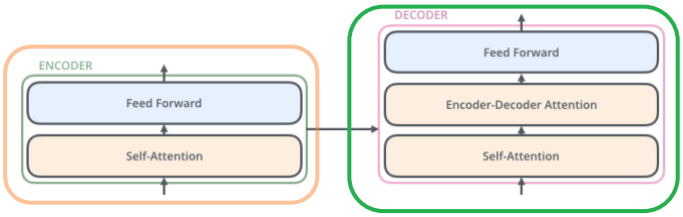



*Figure II.26 Decoder component of the Transformer*[142]

The decoder has both those layers, but between them is an attention layer that helps the decoder focus on relevant parts of the input sentence.



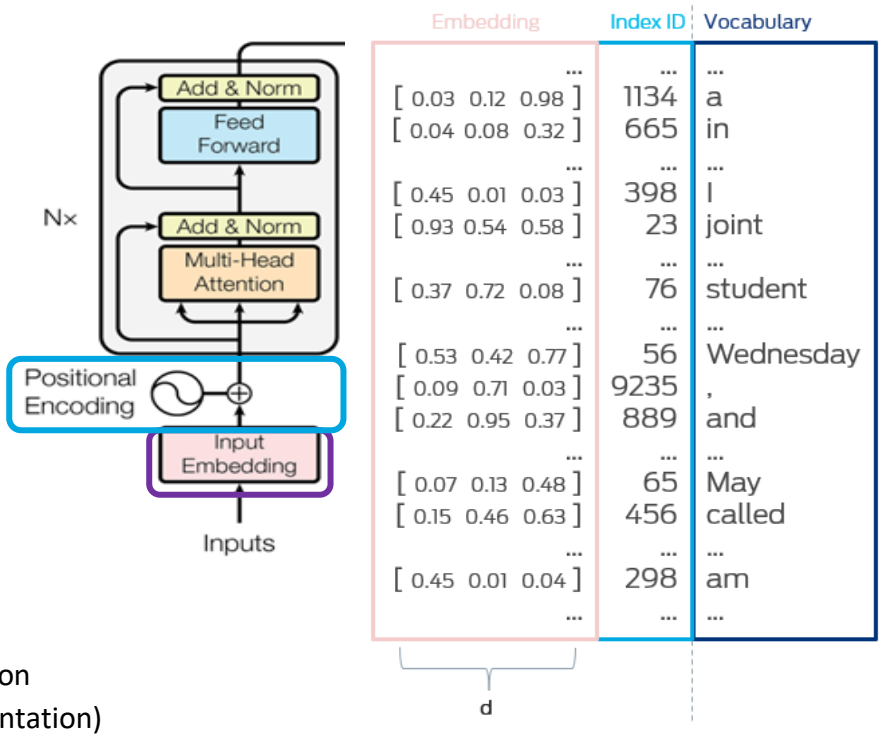

To understand how it operates, we will walk through the Transformer architecture.

### *The embedding algorithm*

*Figure II.27 Mapping of the words, their matching index ID and Embeddings*[142]

The first step of modeling with Transformer is the transformation of inputs (words) into embeddings.
Each input word is turned into a vector using an embedding algorithm as seen in paragraph II.2.2.4. Embedding is used to map words or phrases from a vocabulary to a corresponding vector of real numbers. This vector representation has two important and advantageous properties:

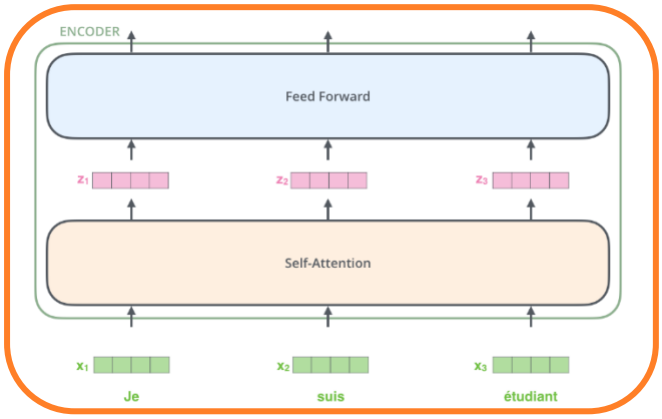
- Dimensionality Reduction (more efficient representation)
- Contextual Similarity
- (more expressive representation)



Since the model has been trained from a large corpus of unique words. Each of these words have a unique ID, known as vocabulary index. Therefore, the embedding algorithm converts the input word into its corresponding word embedding. The embedding only happens in the bottom-most encoder. The input goes through a positional encoding which allows to know the place of each word in the text. Important for context.

*Figure II.28 Flow of the words in the bottom encoder*[142]

After embedding, each of the words flows through each of the two layers of the encoder. The key property of the Transformer is that the word in each position flows through its own path in the encoder. There are dependencies between these paths in the self-attention layer. However, the feed-forward layer does not have those dependencies, and thus the various paths can be executed in parallel while flowing through the feed-forward layer.




### *Self-attention in detail*

As the model processes each word (each position in the input sequence), self- attention allows it to look at other positions in the input sequence for clues that can help lead to a better encoding for this word.

*Figure II.29 Weights and Query/key/value matrix for each word*[142]

The first step in calculating self-attention is to create three vectors from each of the encoder's input vectors (in this case, the embedding of each word). For each word, a Query vector, a Key vector, and a Value vector is created by multiplying the embedding by three matrices that are trained during the training process. Multiplying $X_1$ by the $W^Q$ weight matrix produces $q_1$. The weights are already learned when the model has been trained on large amount of data.

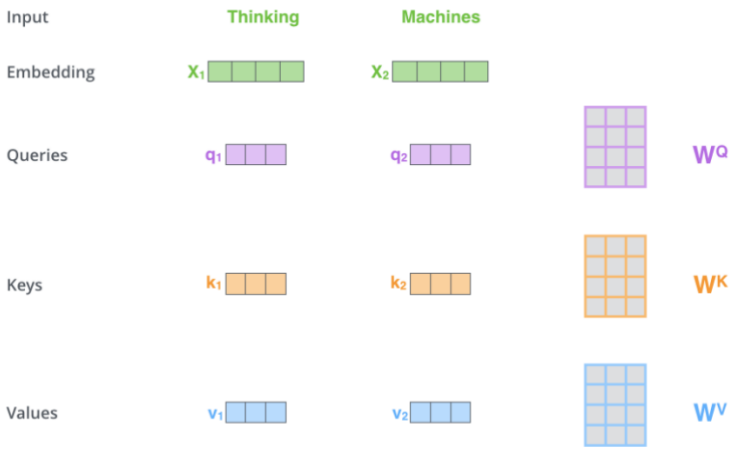


*Figure II.30 Illustration of the dot product of query and key vectors*[142]

The second step in calculating self-attention is to calculate a score which determines how much focus to place on other parts of the input sentence. It is calculated by taking the dot product of the query vector with the key vector. The score of the word in the first position is the dot product of $q_1$ and $k_1$. The second score would be the dot product of $q_1$ and $k_2$.

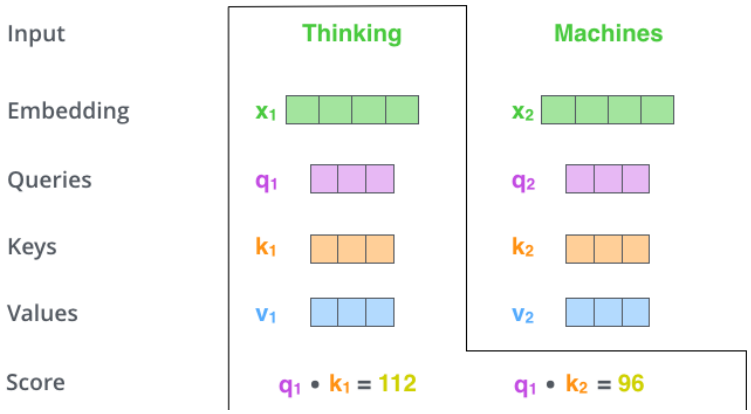
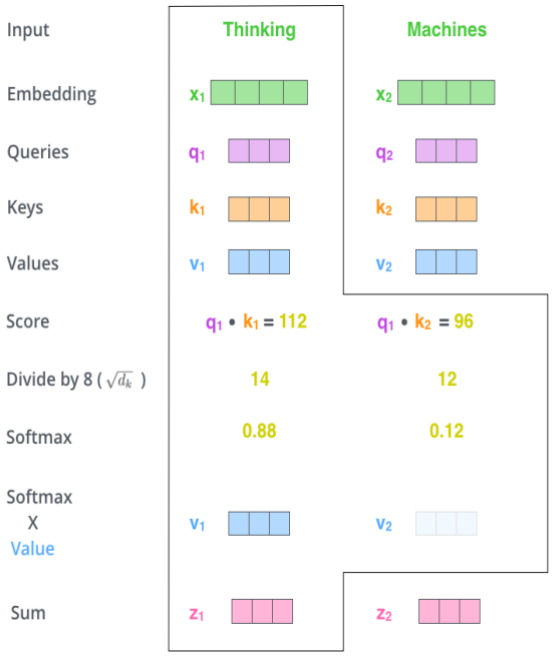



*Figure II.31* Illustration of the production of the output of the self-attention layer[142]




The third and fourth steps are to divide the scores by the square root of the dimension of the key vectors (64 for the vanilla Transformer), leading more stable gradients. The result is then passed through a softmax operation leading to positive scores add up to 1. This softmax score determines how much each word will be expressed at this position.

The fifth step is to multiply each value vector by the softmax score (in preparation to sum them up). The intuition here is to keep intact the values of the word(s) we want to focus on and drown-out irrelevant words.

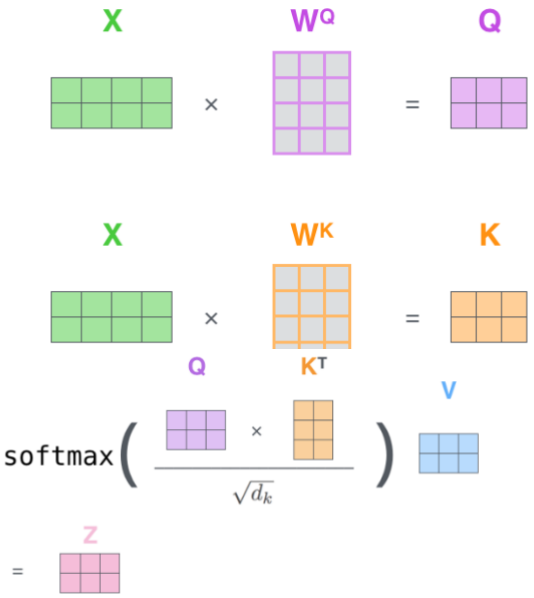
The sixth step is to sum up the weighted value vectors. This produces the output of the self-attention layer at this position (for the first word). The resulting vector can be sent along to FFNN.

### *Matrix Calculation of Self-Attention*

The first step is to calculate the Query, Key, and Value matrices. To do so, the embeddings are packed into a matrix X which is multiplied by the trained weight matrices trained ($W^Q, W^K, W^v$). Every row in the X matrix corresponds to a word in the input sentence.

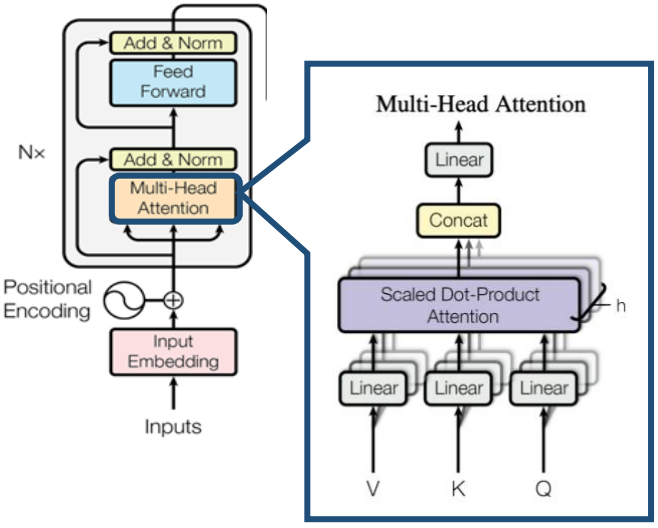
Finally, step two to six can be condensed in one formula to calculate the outputs of the self-attention layer.

*Figure II.32 Illustration of the matrix calculation of self-attention*[142]




### *Multi-head attention*

*Figure II.33 Multi-head attention of the Transformer*[66]

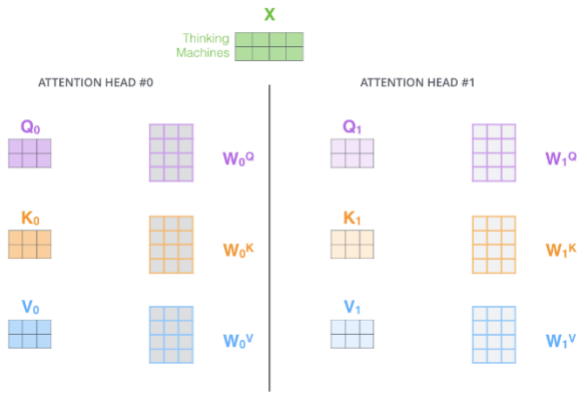



Multi-head attention is the most important module of the whole architecture, it defines the essence of the transformer. This improves the performance of the attention layer in two ways:

- It expands the model's ability to focus on different positions. As seen on Figure II.33, $z_1$ contains a little bit of every other encoding, but it could be dominated by the actual word itself.
- It gives the attention layer multiple "representation subspaces". With multi-headed attention we have not only one, but multiple sets of Query/Key/Value weight matrices (the vanilla Transformer uses 8 attention heads, resulting in 8 sets for each encoder/decoder). Each of these sets is randomly initialized. Then, after training, each set is used to project the input embeddings (or vectors from lower encoders/decoders) into a different representation subspace.

*Figure II.34 Illustration of the separated weight matrices in Transformer*[142]

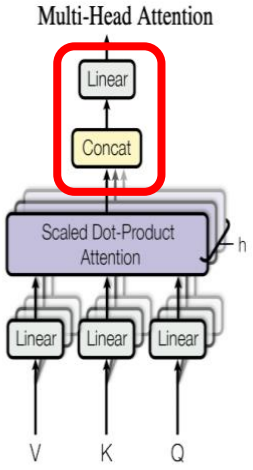
With multi-headed attention, the Q/K/V weight matrices are maintained separated for each head resulting in different Q/K/V matrices. As we did before, we multiply $X$ by the $QW_i^Q, KW_i^K, VW_i^V$ matrices to produce $(Q, K, V)$ matrices.




$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$
$$\text{Where head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

*Figure II.35 Concatenation and linear normalization layers of the Transformer*[66]

The feed-forward layer is designed to receive a single matrix (a vector for each word) as input, not eight separate matrices. To accommodate this, the matrices of the attention heads are concatenated, resulting in a combined matrix. This matrix is then multiplied by a weight matrix $W^0$, which was trained alongside the model. The resulting $z$ matrix contains information from all the attention heads and is passed on to the FFNN for further processing.




*Summary of the attention process*

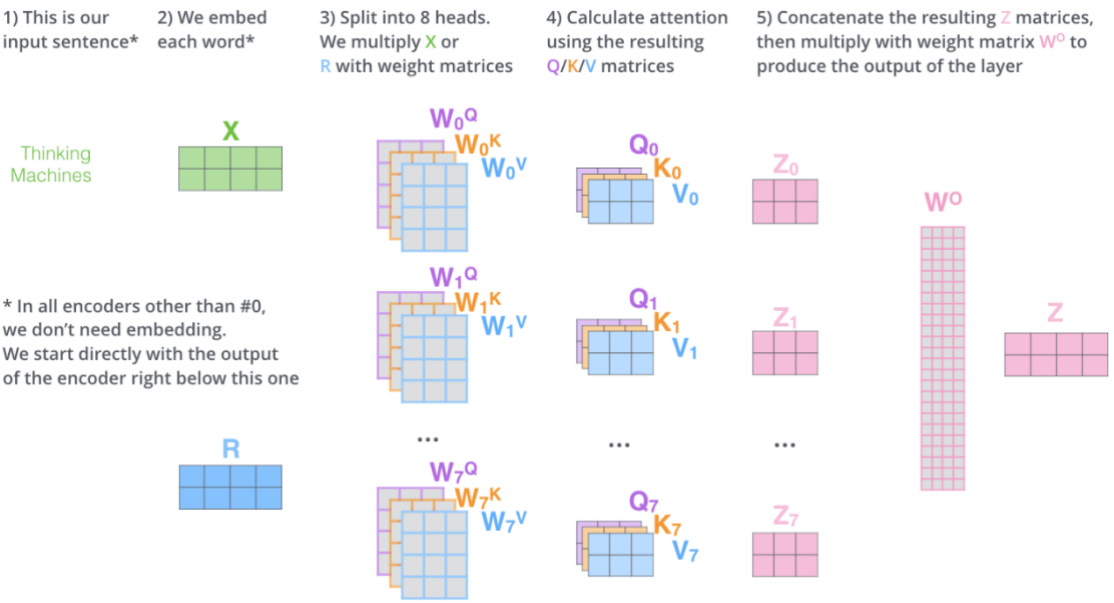



Figure II.36. Illustration of the summary of the attention process[142]

*Figure II.37 Encoder-decoder attention layer of the Transformer*[66]

The encoder starts by processing the input sequence. The output of the top encoder is then transformed into a set of attention vectors K and V. These are used by each decoder in its "encoder-decoder attention" layer as seen on Figure II.37. which helps the decoder focus on appropriate places in the input sequence. After finishing the encoding phase, we begin the decoding phase. Each step in the decoding phase outputs an element from the output sequence.

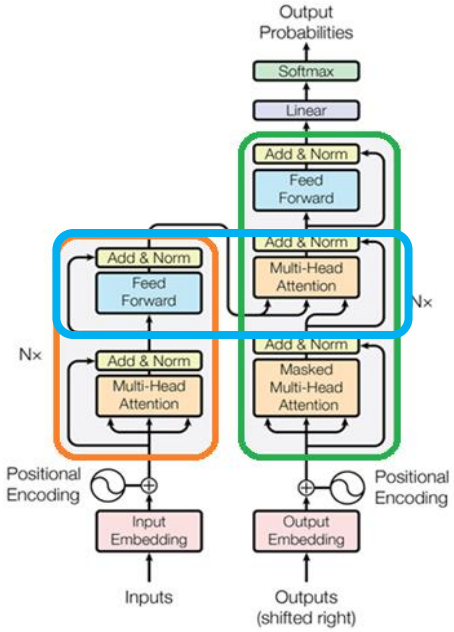
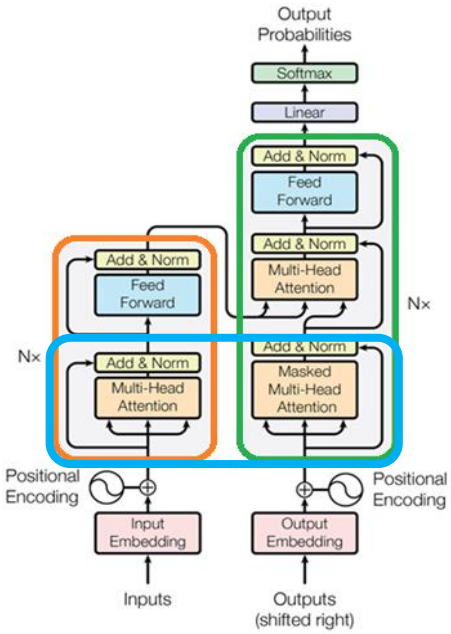



*Figure II.38 Bottom decoder layer of the Transformer*[66]

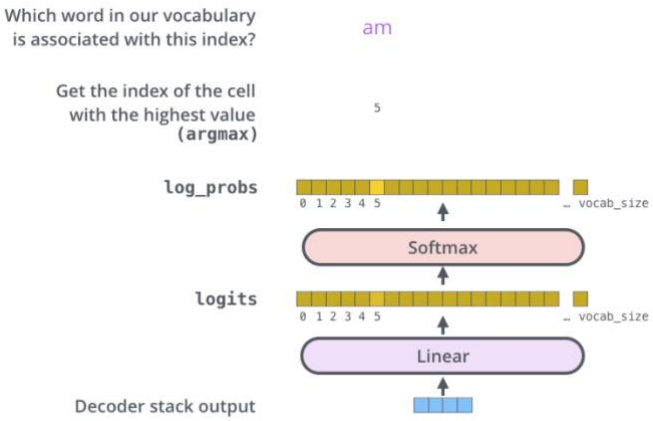



The following steps repeat the process until a special symbol is reached indicating the transformer decoder has completed its output. The output of each step is fed to the bottom decoder in the next time step, and the decoders bubble up their decoding results just like the encoders did. And just like we did with the encoder inputs, we embed and add positional encoding to those decoder inputs to indicate the position of each word. The self-attention layers in the decoder operate in a slightly different way than the one in the encoder, in fact, in the decoder, the self-attention layer is only allowed to attend to earlier positions in the output sequence. This is done by masking future positions (setting them to -inf) before the softmax step in the self-attention calculation. The "Encoder-Decoder Attention" layer works just like multiheaded self-attention, except it creates its Queries matrix from the layer below it and takes the Keys and Values matrix from the output of the encoder stack.

### *The Final Linear and Softmax Layer*

*Figure II.39 Illustration of the final linear and softmax layer*[142]

The final linear layer followed by a softmax layer turns the output vectors to words. The linear layer is a simple fully connected neural network that projects the vector produced by the stack of decoders, into a larger vector called a logits vector. The softmax layer then turns these scores into probabilities (all positive, all add up to 1.0). The cell with the highest probability is chosen, and the word associated with it is produced as the output for this time step.




## Transformers Unsupervised Pre-Training

The main strength of language models based on Transformers, that powered them to reach states of the art performance on various NLP tasks, is inherited from their unsupervised pre-training step. Unsupervised pre-training refers to the step in which language models learn a contextual representation from large unlabeled data prior to a task-specific training step.
Once pre-trained, these models are named pre-trained foundation models[143] (PFMs).

When pretraining techniques are applied to the NLP domain, well-trained language models can capture rich knowledge beneficial for downstream tasks, such as long-term dependencies, hierarchical relationships, etc.
Early pretraining is a static technique, such as Word2vec[105], but static methods were difficult to adapt to different semantic environments. Therefore, dynamic pretraining techniques are proposed, such as BERT[67] and XLNet[144].

### *Training hyperparameters*

Several training hyperparameters are to be defined for the pre-training step, as most of them are also involved in supervised training.

- Epoch: refers to one cycle through the full training dataset.
- Iteration: number of batches or steps through partitioned packets of the training data, needed to complete one epoch.
- Batch size: number of samples propagated through the network.
- Learning rate: parameter that scales the magnitude of the model's weight updates in order to minimize the network's loss function.
- Optimizer: parameter managing the weights or learning rate variations.

All these hyperparameters are intricated, and optimizing hyperparameters with LLM is not straightforward.

The learning rate controls how quickly the model is adapted to the task. Smaller learning rates require more training epochs given the smaller changes made to the weights each update, whereas larger learning rates result in rapid changes and require fewer training epochs. As found in[145,146], larger models can typically use a larger batch size, but require a smaller learning rate.
Regarding the optimizer, stochastic gradient descent (SGD) is still the most used algorithm in deep learning. However, adaptive methods like Adam[147] have been observed to outperform SGD across important tasks, such as attention models[148].

#### *Unsupervised pre-training advantages*

Erhan et al.[149] have demonstrated that pre-training acts as a regularization scheme, enabling better generalization in deep neural networks. Unsupervised pre-training provides an unusual form of regularization: minimizing variance and introducing bias towards configurations of the parameter space that are useful for unsupervised learning.
The regularization effect is a consequence of the pre-training procedure establishing an initialization point of the fine-tuning procedure inside a region of parameter space in which the parameters are henceforth restricted. The parameters are restricted to a relatively small volume of parameter space that is delineated by the boundary of the local basin of attraction of the supervised fine-tuning cost function.
Unsupervised pre-training also mitigates problems with data scarcity by pre-training over a large, diverse dataset. Indeed, training data can be derived from any unlabeled text corpus, that is, there is an unlimited amount of training data in the pretraining process. Therefore, data augmentation use is not always useful when using Transformer models.

### Embeddings and pre-training corpora for clinical notes

The choice of the embeddings used for clinical NLP tasks, thus the corpora on which they have been trained on can have an impact on downstream tasks. Embeddings are a useful and versatile tool with the ability to perform well in many predictive tasks. Their utility extends to (often noisy) clinical note data. Embeddings trained on domain specific corpora can capture valuable information contained in clinical free-text, at relatively low cost (without the need for manual annotation or expert curation). As an example, Wang et al. compared word embeddings from several corpora before performing clinical information extraction, biomedical information retrieval, and relation extraction[150]. They found that word embeddings trained from clinical notes and literature capture word semantics better than more general corpora such as GloVe or Google News. Huang et al.[151] tested their ClinicalBERT (embeddings trained on MIMIC-III[152], a 2 million clinical notes dataset) model on downstream tasks such as predicting hospital readmission within 30 days using the MIMIC-III dataset. They used both discharge summaries and early clinical notes for hospital readmission prediction. ClinicalBERT outperformed the state-of-the-art in both cases. Alsentzer et al.[153] evaluated ClinicalBERT model on two NER tasks (concept extraction and entity extraction), and a natural language inference task; they showed that Clinical BERT performed better than the original BERT and BioBERT on NER and NLI tasks.
To acquire embeddings, massive volumes of domain-specific corpus are required, as explained in Dieng et al. work[154], and one of the main issues for clinical data is that data privacy policies

often prevent the release of any models learned from those data. The result is that researchers are forced to create their own models using their own data which leads to difficulties in reproducibility. However, recently, great efforts have also been made for providing French clinical annotated corpora such as MERLOT (Medical Entity and Relation LIMSI annOtated Text corpus)[155], CAS (Corpus of Clinical Case)[156] or CLISTER (Corpus for Semantic Textual Similarity in French Clinical Narrative)[157].

#### *Unsupervised pre-training drawback*

The large amount of scrapped data used for unsupervised pre-training has drawbacks:

- Compute Requirements: Pre-training large language models induces computational costs in terms of time and memory. However, the supervised fine-tuning step of these models is quicker leading to a faster convergence towards better accuracy.

- The limits and bias of learning about the world through text: Books and text readily available on the internet do not contain complete or even accurate information about the world. Recent work has shown that certain kinds of information are difficult to learn via just text and other work has shown that models learn and exploit biases in data distributions.

- Still brittle generalization: Although most approaches improve performance across a broad range of tasks, current deep learning NLP models still exhibit surprising and counterintuitive behavior - especially when evaluated in a systematic, adversarial, or out-of-distribution way.

#### *Pre-training tasks*

Several pre-training tasks can be performed such as:

- predict the next word (GPTs)
- predict a mask of a token (BERT)
- predict a mask of several tokens (SpanBERT)
- change the order of the tokens and find those that have been changed (ELECTRA)
- training with several languages (XLM)
- training with corruptions (BART, T5)

### Different types of Transformers

Unlike previous approaches that use convolutional and recurrent modules to extract features, BERT learns bidirectional encoder representations from transformers trained on large datasets as contextual language models. Similarly, the Generative Pretrained Transformer (GPT) method uses transformers as feature extractors and is trained on large datasets using an autoregressive paradigm.

Transformer models differ in their training strategies, model architectures, and use cases. They can be categorized into two types: encoder-decoder or encoder-only language models and decoder-only language models. The Table II.1 summarizes the characteristics and the representative LLMs of each type available in 2023.

| | Characteristics | | LLMs |
|---|---|---|---|
| Encoder-Decoder or Encoder-only (BERT-style) | Training:<br>Model type:<br>Pre-train task: | Masked Language Models<br>Discriminative<br>Predict masked words | ELMo[111], BERT[67], RoBERTa[70], DistilBERT[113], BioBERT[158], XLM[159,] Xlnet, ALBERT[160], ELECTRA[114], T5[161], GLM, XLM-E[162], ST-MoE[163,164], AlexaTM[163] |
| Decoder-only (GPT-style) | Training:<br>Model type:<br>Pre-train task: | Autoregressive Language Models<br>Generative<br>Predict next word | GPT-3[165], OPT[166], PaLM[167], BLOOM[168], MT-NLG[169], GLaM[170], Gopher[171], chinchilla[172], LaMDA[173], GPT-J[174], LLaMA[175], GPT-4, BloombergGPT[176], PALM-E[177] |

*Table II.1* Summary of Large Language Models (inspired from[178]). Models are chronologically presented in the table.

### *BERT*

Developed in 2018 by researchers at Google AI Language, BERT (Bi-directional Encoder Representations from Transformers)[67] is a transformer decoder-only model that predicts which words are masked and determine whether two sentences are contextual.

#### *BERT architecture*

Let's define:

- Parameters: Number of learnable variables/values available for the model
- Transformer Layers: Number of Transformer blocks. A transformer block transforms a sequence of word representations to a sequence of contextualized words (numbered representations).
- Hidden Size: Layers of mathematical functions, located between the input and output, that assign weights (to words) to produce a desired result.
- Attention Heads: the size of a transformer block
- Processing: Type of processing unit used to train the model
- Length of Training: Time it took to train the model

When BERT was introduced in "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"[67], 2 model sizes were available: BERTbase and BERTlarge as seen on with the sizes and architectures depicted in Table II.2 BERTbase and BERTlarge size and architecture

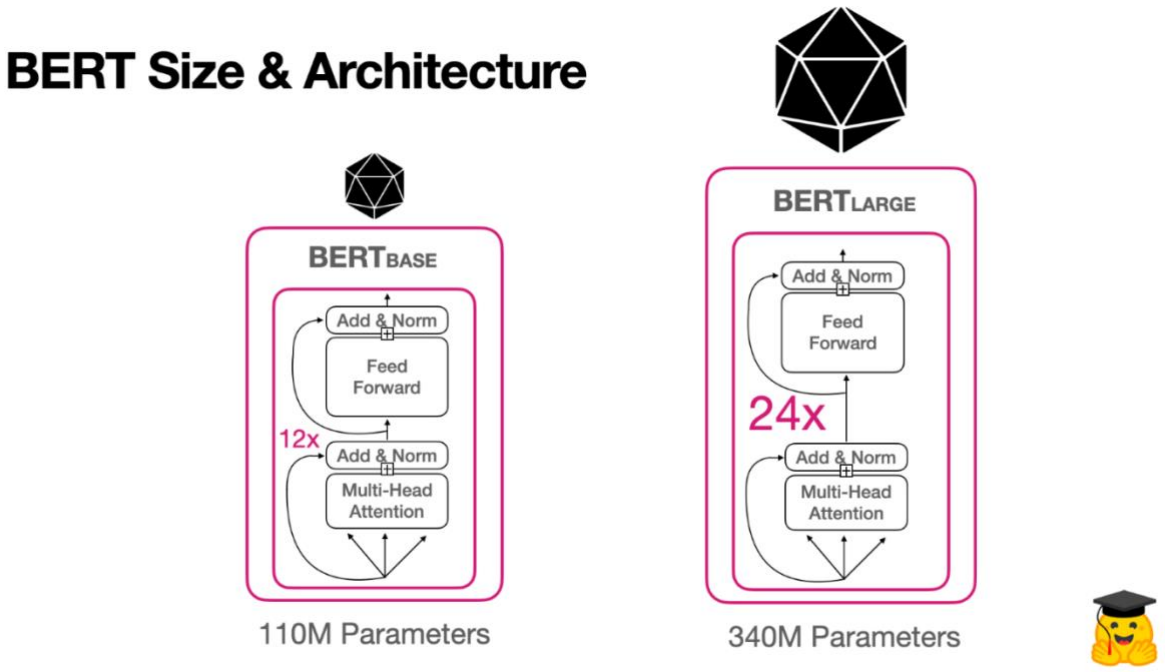



*Figure II.40* BERTbase and BERTlarge size and architecture illustration (from Hugging face blog[179])

| | Transformer Layers | Hidden Size | Attention Heads | Parameters | Processing | Length of Training |
|---|---|---|---|---|---|---|
| BERTbase | 12 | 768 | 12 | 110M | 4 TPUs | 4 days |
| BERTlarge | 24 | 1024 | 16 | 340M | 16 TPUs | 4 days |

*Table II.2* BERTbase and BERTlarge size and architecture

BERT is a trained Transformer Encoder stack. Both BERT model sizes have a large number of encoder layers, 12 for BERTbase and 24 for BERTlarge. These also have larger feedforward-networks (768 and 1024 hidden units respectively), and more attention heads (12 and 16 respectively) than the default configuration in the reference implementation of the Transformer in the initial paper (6 encoder layers, 512 hidden units, and 8 attention heads) as seen on Figure II.40.

BERT uses bidirectional self-attention which allows each token's representation to be adapted based on all other tokens within a sequence.

### *BERT pre-training*

WordPiece embeddings[180] with a 30,000 tokens as seen in section II.2.2.1 were used for pre-training BERT. The input representation is able to represent both a single and a pair of sentences in one token sequence. The first token of every sequence is always a special classification token [CLS]. The final hidden state corresponding to this token is used for the

classification task. The two sentences are separated using the [SEP] token. In the case of sentence pair, a segment embedding is added which indicates whether the token belongs to sentence A or sentence B as seen on Figure II.41.

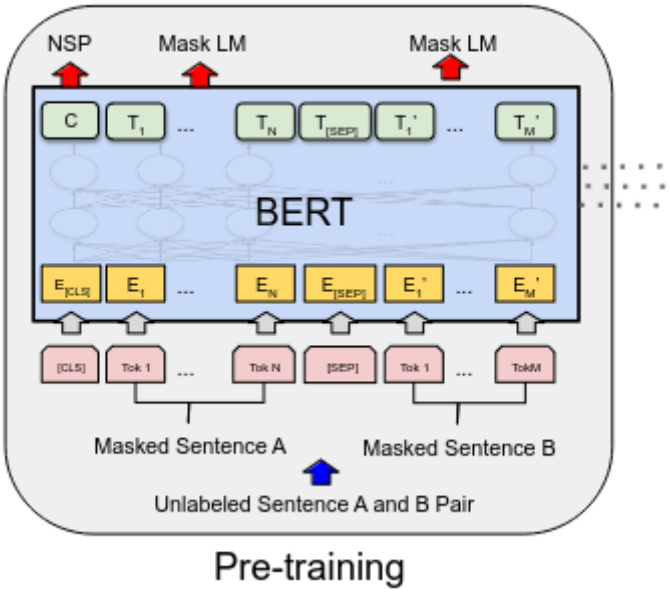



*Figure II.41* Pre-training procedure of BERT. (from [67])

For a given token, its input representation is constructed by adding the corresponding token, segment and position embedding. A visualization of this construction can be seen in Figure II.42.

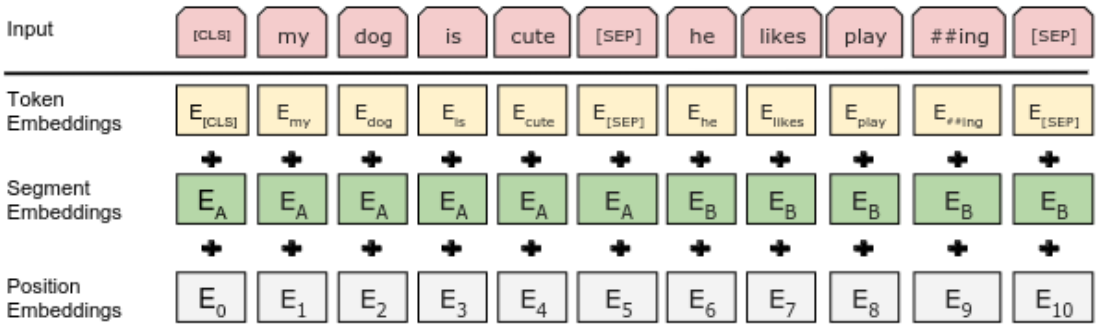



*Figure II.42* BERT input representation (from [67])

BERT was trained on a large corpus of English texts (Wikipedia and BookCorpus) in a self-supervised manner.

Wikipedia is a large collection of 4.4 million articles that contains 1.9 billion words in its 2022 version (21.26 GB). Devlin et al.[67] used a filtered version of this corpus by extracting only the text passages and ignoring lists, tables, and headers, resulting in a 2.5 million words.

BookCorpus is a large collection of free novel books written by unpublished authors, which contains 11,038 books (around 74M sentences and 1G words) of 16 different sub-genres (e.g., Romance, Historical, Adventure, etc.) which represents a dataset of 4.85 GB[181]

BERT was pre-trained using two different tasks:

- Masked language modeling: the model randomly masks 15% of the words in the input, sequence by sequence, then run the entire masked sentence through the model and must predict the masked words. It allows the model to learn a bidirectional representation of the sentence.
- Next sentence prediction: the models concatenates two masked sentences as inputs during pretraining. Sometimes they correspond to sentences that were next to each other in the original text, sometimes not. The model then has to predict if the two sentences were following each other or not.

The model learns an inner representation of the English language that can then be used to extract features useful for downstream task such as classification or named entity recognition.

### *BERT fine-tuning*

Fine-tuning is an approach to transfer learning in which the weights of a pre-trained model are trained on new data. Fine-tuning can be done on the entire artificial neural network or on only a subset of its layers, in which case the layers that are not being fine-tuned are "frozen". Models such as BERT are usually fine-tuned by reusing the model's parameters as a starting point and adding a task-specific layer trained from scratch[182]. Fine-tuning BERT is straightforward since the self-attention mechanism in the Transformer allows BERT to model many downstream tasks (whether they involve single text or text pairs) by swapping out the appropriate inputs and outputs. For each task, the task-specific inputs and outputs are plugged into BERT and fine-tune all the parameters end-to-end.

### *BERT models and performances for clinical NLP tasks*

Since its creation, several models based on the BERT architecture have been proposed, some of which are dedicated to health data. As of May 2023, 202,934 BERT models were available in the Hugging Face library. BERT can be:

- pre-trained on a various number of languages;
- multilingual;
- cased: the text is the same as the input text (no changes) and the accent marks are preserved;
- uncased: the text is lower cases and the accent marks will be removed prior to the WordPiece tokenization step
- fine-tuned on domain specific corpora
- fine-tuned for specific tasks

Several BERT models have been proposed for different NLP tasks specific to health data, such as BioBERT[158], MedBERT[183], BEHRT[184,185], G-BERT, etc. Most of these models are trained on

English biomedical language, but recently DrBERT (a French biomedical model) has been published and has shown an improvement in performance on most tasks compared to previous techniques. In fact, DrBERT was evaluated on 11 different practical biomedical applications for French, including named entity recognition (NER), part-of-speech tagging (POS), binary/multi-class/multi-label classification, and multiple-choice question answering. The results showed that the from-scratch pre-trained strategy is still the most effective for BERT language models on French biomedical text[186]. Regarding ED EHR clinical notes, Valmianski et al[187] compared 6 different BERT based models for a multi-class classification task on the ED chief of complaint. They used BERT-base, BioBERT, ClinicalBERT, they also pre-trained ClinicalBERT on patient progress notes (from encounters unrelated to the ED) and added another step of pre-training with a chief of complaint dataset. They also used TF-IDF embeddings with the last model as a baseline. They found that the model trained on both progress notes and chief of complaint clinical notes had better performance.

### *GPTs*

#### *GPT architecture*

Language models such as GPT-3[165] have revolutionized modern deep learning applications for NLP. Interestingly, however, most of the technical novelties of GPT-3 were inherited from its predecessors GPT and GPT-2[188,189]. As such, an understanding of GPT and GPT-2 is useful.
Both GPT and GPT-2 use a decoder-only transformer architecture. Therefore, the entire encoder and encoder-decoder self-attention blocks in the decoder are absent from GPT architecture. Each layer of the decoder consists of a masked self-attention layer followed by a feed forward neural network as can be seen on Figure II.43. Using masked self-attention yields an autoregressive architecture (i.e., meaning that the model's output at time t is used as input at time t+1) that can continually predict the next token in a sequence.

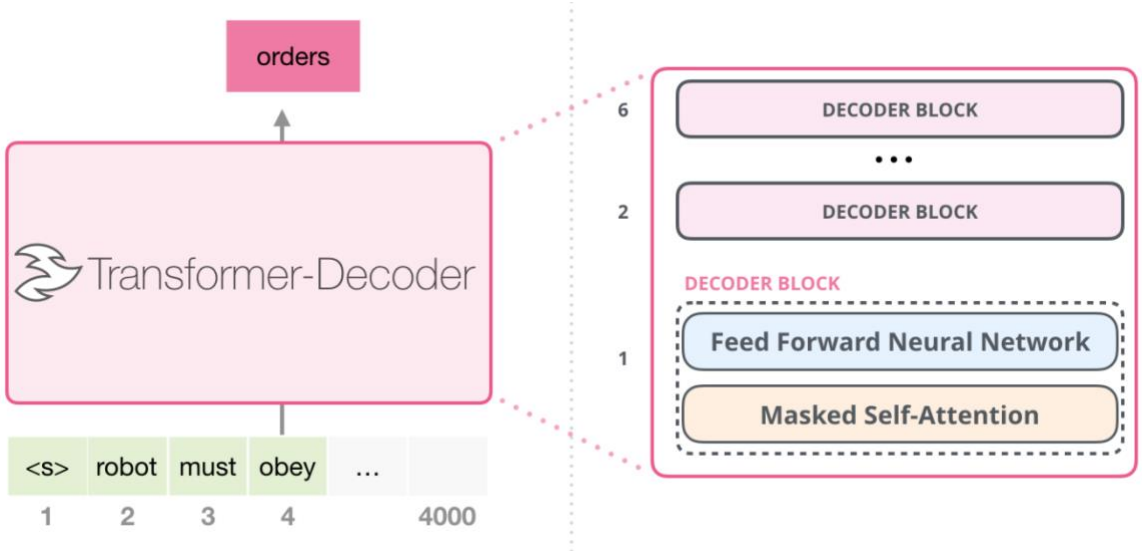



*Figure II.43* Decoder blocks of a decoder-only Transformer. The first decoder block Is expanded. (From[142])

Original GPT architecture

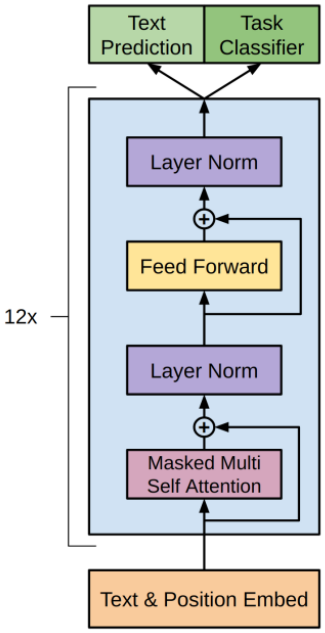



*Figure II.44 GPT architecture and training objectives (From[188])*

Introduced in 2018 by Radford et al.[188], GPT uses a 12-layer, decoder-only transformer architecture that matches the original transformer decoder[66] (aside from using learnable positional embeddings). This model was first proposed by Liu et al[190] and had several purposes:

- Reduce model parameters for a given hyper-parameter set
- Reduce error propagation from both input and output time-steps during training (since the vanilla Transformer is forced to predict the next token in the input as well as the output)
- Limit redundant information re-learning about language in the encoder and decoder when monolingual tasks text-to-text are performed
- Ease the optimization.

GPT applies a multi-headed self-attention operation over the input context tokens followed by position-wise feedforward layers to produce an output distribution over target tokens[188]:

$$h_0 = UW_e + W_p$$
$$h_1 = \text{transformerblock}(h_{l-1}) \forall_{i_i} \in [1, n] \quad (II.29)$$

$$P(u) = \text{softmax}\,(h_n W_e^T)$$

Where $U = (u_{-k}, \dots, u_{-1})$ is the context vector of tokens, $n$ is the number of layers, $W_e$ is the token embedding matrix, and $W_p$ is the position embedding matrix.

GPT has 12 layers with masked self-attention heads (768 dimensional states and 12 attention heads). For the position-wise feed-forward networks, 3072 dimensional inner states were used. The optimizer was Adam[147] with a maximum learning rate of 2.5e-4.

The tokenizer used was the BPE as described in the section Byte-Pair Encoding, as used in GPT section. The vocabulary was set to with 40,000 merges[108]. Residual, embedding, and attention dropouts with a rate of 0.1 for regularization were used. The activation function used was the Gaussian Error Linear Unit. Just like BERT, GPT was pre-trained with the BooksCorpus[181] corpora.

### GPT-2 architecture

Released in 2019 by Radford et al.[189], GPT-2 model largely follows the details of the GPT model with a few modifications:

- Expansion of the vocabulary to 50,257 (40,000 for GPT) merges
- Increase of the context size (from 512 to 1024)
- Moving of the layer normalization to the input of each sub-block
- Addition of a layer after the final self-attention block

The initial GPT-2 model had several sizes going from 12 decoder blocks and 768 dimensions to 48 decoder blocks and 1600 dimensions. GPT-2 small had 117 million parameters while GPT-2 had 1542 million parameters making it the largest language model in 2019.

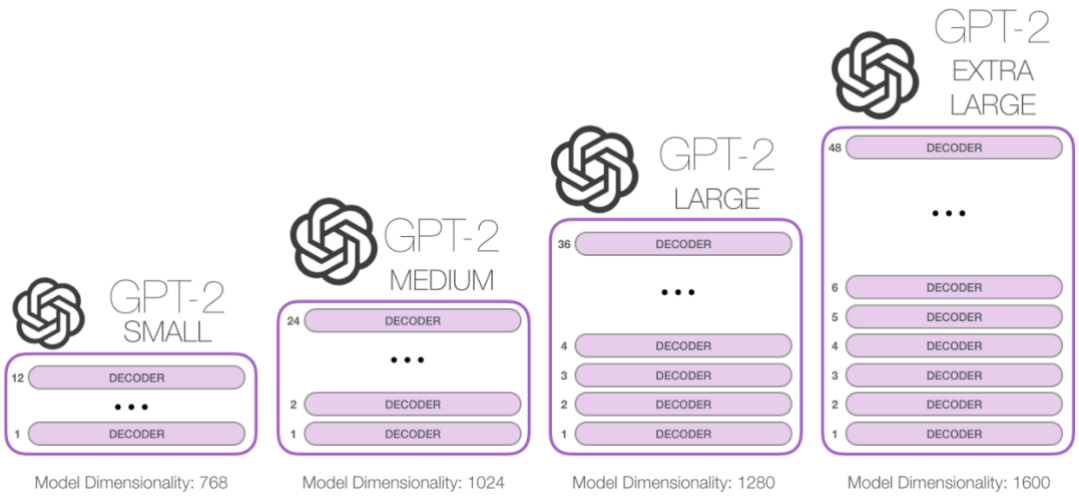



*Figure II.45* GPT-2 sizes and dimensionality (from[142])

The model was pre-trained with WebText, a dataset containing 8 million documents representing 40GB of text. The authors scraped web pages solely curated/filtered by humans. The model is pre-trained using a language modeling objective, but it performs no fine-tuning, choosing to solve downstream tasks in a zero-shot manner instead. Put simply, GPT-2 performs multi-task learning by:

- Pre-training a generic LM over raw textual data
- Using textual "prompts" to perform zero-shot inference on a variety of tasks

#### *GPT models and performances*

Decoder-only models such as GPT have gradually come to dominate the development of LLMs. In the early stages of LLM development, decoder-only models were not as popular as encoder-only (BERT) and encoder-decoder models. However, after 2021, with the introduction of the

game-changing LLMs - GPT-3 - decoder-only models experienced a significant boost. Meanwhile, encoder-only models began to fade after the initial explosive growth brought about by BERT.
An example of pre-trained foundation model application is ChatGPT. ChatGPT is fine-tuned from the generative pretrained transformer GPT-3.5. ChatGPT applies reinforcement learning from human feedback[191], which has become a promising way to align LLMs with humans' intent[192].
However, when EHR clinical notes are concerned, studies about the use of GPT for information extraction is sparse.

### II.2.6 Data Augmentation in NLP

When dealing with classification for medical data, datasets are usually imbalanced[193–195]. This issue has been addressed with several techniques such as re-sampling methods or data augmentation at several steps of the pipeline. Data augmentation (DA) refers to methods used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data. Such methods alleviate data scarcity scenarios where deep learning techniques may fail.
DA methods can be framed into three categories, including paraphrasing, noising, and sampling:

#### *II.2.6.1. Paraphrasing*

The paraphrasing-based methods generate augmented data that has limited semantic difference from the original data, based on proper and restrained changes to sentences. The augmented data convey very similar information as the original form.
Paraphrasing consists of several levels, including lexical paraphrasing (word level), phrase paraphrase, and sentence paraphrase.

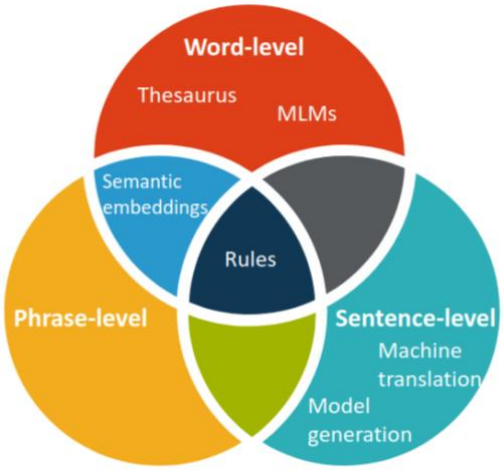



*Figure II.46* Methods of paraphrasing in text data augmentation

**Word-level:**

- Synonym or hypernyms replacement with thesauri: Words can be replaced in the original text with their synonyms and hypernyms to obtain a new way of expression

while keeping the semantics of the original text as unchanged as possible[196]. Thesauri like WordNet[197] contain such lexical triplets of words and are often used as external resources. Thesauri are easy to use but the scope and Part-Of-Speech (POS) of augmented words are limited. The sentence semantics can also be affected if there are too many substitutions.

- Semantic embedding replacement: This method overcomes the limitations of replacement range and parts of speech in the thesaurus-based method. It uses pre-trained word embeddings, such as Glove[106], Word2Vec[105], FastText[198], etc., and replaces the original word in the sentence with its closest neighbor in embedding space. Semantic embeddings are easy to use and have a higher replacement hit rate and a more comprehensive range. However, this method cannot resolve the ambiguity problem (see section II).

- Language models: Pretrained language models have become mainstream models in recent years due to their excellent performance. Masked language models (MLMs) such as BERT[67] and RoBERTa[70] can predict masked words in text based on context, which can be used for text data augmentation. Moreover, this approach alleviates the ambiguity problem since MLMs consider the whole context.

**Sentence-level:**

- Machine translation: translation is a natural means of paraphrasing. With the development of machine translation models and the availability of online APIs, machine translation is popular as an augmentation method in many tasks. Back translation is a method where the original text is translated into other languages, and then translated back to obtain the augmented text in the original language. Different from word-level methods, back-translation does not directly replace individual words but rewrites the whole sentence in a generated way. Unidirectional translation method directly translates the original text into other languages once, without translating it back to the original language. This method usually occurs in a multilingual scene. These methods have a wide range of applications and guarantees of correctness of syntax and unchanged semantics but there is a poor controllability and limited diversity because of the fixed machine translation models.

- Model generation: Some methods employ Seq2Seq, Gan or Transformers models to generate paraphrases directly. Such models output more diverse sentences given proper training objects. These models have a wide range of applications, but they require training data and have a high computing cost.

**All levels:**

Rules: To ensure the preservation of sentence meaning, a particular approach in NLP requires the use of heuristics. Some studies have utilized existing dictionaries or fixed heuristics to generate word-level and phrase-level paraphrases while employing regular expressions to modify the form of the text without altering the semantics. This includes the use of abbreviations and verb prototypes, as well as the handling of modal verbs and negation. Other studies have generated sentence-level paraphrases for the

original sentences by utilizing rules based on dependency trees. This involves rotating the target fragment around the root of the dependency parse structure, which does not negatively affect the original meaning of the sentence. Although this rule-based method preserves the original sentence semantics, it requires artificial heuristics and has a limited range of coverage and variation.

#### *II.2.6.2. Noising*

Noising-based methods add more continuous or discrete noises to the original data and involve more changes. The noising-based methods add faint noise that does not seriously affect the semantics, to make it appropriately deviate from the original data. Humans greatly reduce the impact of weak noise on semantic understanding through their grasp of linguistic phenomena and prior knowledge, but this noise can pose challenges for models. Thus, this method not only expands the amount of training data but also improves model robustness.

Noising can be performed with[199] :

- Random Insertion: selection of a random (non stopword) word in a sentence, selection of a random synonym of this word and random insertion of it, n times[200]
- Random Swap: Random selection of two words in the sentence and swap their positions; n times.
- Random Deletion: Randomly remove each word in the sentence with probability p.
- Random Substitution: Randomly replace words or sentences with other strings. Different from the above paraphrasing methods, this method usually avoids using strings that are semantically similar to the original data.

#### *II.2.6.3. Sampling*

Sampling-based methods master the distribution of the original data to sample new data as augmented data. The sampling-based methods are task-specific and require task information like labels. Such methods not only ensure validity but also increase diversity.

- Non-pre-trained models: Train a target-to-source model and use the model to generate source sentences from target sentences by learning the internal mapping between the distributions of the target and the source[201].

- Pre-trained models: Use of pre-trained Transformers for data augmentation with dedicated models such as LAMBADA[202] (GPT2-based) or SSMBA[203] (BERT-based) or adapted generative models mostly based on GPT2[204–206]
- Self-training: Use of fine-tuned Transformers on the original data, then use the model to label unlabeled sentence pairs[207] or use of data distillation into the self-training process[208] or transfer existing models from other tasks to generate pseudo-parallel corpus[209,210]

- Mix-up: Use of virtual embeddings instead of generated natural language form text as augmented samples. The existing data is used as the basis to sample in the virtual vector space, and the sampled data may have different labels than the original data.

Imbalance also has to be taken into account for the choice of the metrics when evaluating models or pipelines for our classification task. Despite numerous studies on medical image data augmentation, EHR data augmentation has not been studied much.

### II.2.7 Language Model Evaluation

Two different approaches can be used to evaluate and compare language models:

- Extrinsic evaluation: This approach involves evaluating the models by employing them in an actual task (such as text generation) and looking at their final loss/accuracy. This is the best option as it's the only way to tangibly see how different models affect the task we're interested in. However, it can be computationally expensive and slow as it requires training a full system.

- Intrinsic evaluation: This approach involves finding some metric to evaluate the language model itself, not taking into account the specific tasks it's going to be used for. Perplexity is an intrinsic evaluation method.

#### *II.2.7.1. Perplexity*

Perplexity (PPL) is one of the most common metrics for evaluating language models. Perplexity is defined as the exponentiated average negative log-likelihood of a sequence. If we have a tokenized sequence $X = (x_0, x_1, \dots, x_t)$ , then the perplexity of $X$ is:

$$\mathrm{PPL}(X) = \exp\left\{-\frac{1}{t}\sum_{i}^{t} \log p_\theta\left(x_i|x_{<i}\right)\right\} \qquad (II.30)$$

where $\log p_\theta\left(x_i|x_{<i}\right)$ is the log-likelihood of the ith token conditioned on the preceding tokens $x_{<i}$ according to the model. Intuitively, it can be thought of as an evaluation of the model's ability to predict uniformly among the set of specified tokens in a corpus. Importantly, this means that the tokenization procedure has a direct impact on a model's perplexity which should always be taken into consideration when comparing different models.

#### *II.2.7.2. Macro Average Precision*

Precision expresses the proportion of units a model classifies as positive that are actually positive. In other words, precision indicates how much one can trust the model when it predicts that a record is classified in a given class.

In the case of a multi-class classification, Macro Average Precision over all classes i can be evaluated by the macro-averaging, where the precision over each i class is first calculated and then the precisions over all n classes are averaged. Macro-averaging methods tend to compute an overall average of different measurements, because the numerators of the macro-average precision and macro-average recall are composed of values in the interval [0,1]. There is no relationship with class size, as classes of different sizes are also weighted in the numerator. This implies that the effect of the larger classes has the same importance as that of the smaller ones[211]. Macro-average precision is equal to True Positive Value. Therefore, each clinical note has the same importance using this measure. Note, TP: True Positives and FP: False Positives.

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \quad (II.31)$$

$$\text{Macro precision} = \frac{\sum_{i=1}^{n} \text{precision}_i}{n} \quad (II.32)$$

$$\text{Macro precision} = i\frac{\sum_{i=1}^{n} \frac{TP_i}{TP_i + FP_i}}{n} \quad (II.33)$$

### *II.2.7.3. Micro Precision*

$$\text{Micro Precision} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \quad (II.34)$$

Since in a multi-class framework all false instances are counted, it turns out that:

$$\sum_{i=1}^{n} FP_i = \sum_{i=1}^{n} FN_i \quad (II.35)$$

Since:

$$\text{Micro Recall} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \quad (II.36)$$

Then:

$$\text{Micro Precision} = \text{Micro Recall}$$

### *II.2.7.4. Micro F1-score*

Micro F1-score is defined as a harmonic mean of precision and recall in binary class problem. To extend F1- measure to multi-class, two types of average, micro-average and macro-average are commonly used. In micro-averaging, the F1-measure is computed globally over all class

decisions, precision and recall being obtained by summing over all individual decisions. Micro-averaged F1-measure gives equal weight to each clinical note and is therefore considered as an average over all the clinical note/category pairs[212]. Let's note that Recall is equal to Sensitivity. The f1-score being defined as the harmonic mean of precision and recall:

$$\text{Score } F_1 = 2.\frac{\text{precision } * \text{ recall}}{\text{precision } + \text{ recall}} \quad (II.37)$$

in a multi-class framework:

$$\text{Micro } F_1 \text{ Score } = \frac{2 * \sum_{i=1}^{n} TP_i}{2 * \sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i + \sum_{i=1}^{n} FN_i} \backslash \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \quad (II.38)$$

Then:

$$\text{Micro Precision} = \text{Micro Recall} = \text{Micro } F_1\text{score}$$

### *II.2.7.5. Macro f1-score*

$$\text{Macro}F_1\text{Score} = 2.\frac{\text{Macro precision } * \text{ recall}}{\text{Macro precision } + \text{ recall}} \quad (II.39)$$

### *II.2.7.6. Top-k accuracy*

The top-k accuracy function is a generalization of the accuracy score. The difference is that a prediction is considered correct as long as the true label is associated with one of the top-k predicted scores. If $\hat{f}_{i,j}$ is the most predicetd class for the ith matching sample with the maximum jth score predicted and that $y_i$ is the true value, then the correct predictions fraction on $n_{samples}$ is defined as below:

$$\text{top}_k\text{accuracy}(y, \hat{f}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} \sum_{j=1}^{k} 1(\widehat{f_{i,j}} = y_i) \quad (II.40)$$

### *II.2.7.7. Matthews Correlation Coefficient (MCC) for Multi-class Classification*

In a binary classification setting, accuracy and F1-score score, although popular, can generate misleading results on imbalanced datasets, because they fail to consider the ratio between positive and negative elements. The Matthews correlation coefficient (MCC) can solve this issue, through its mathematical properties that incorporate the dataset imbalance and its invariantness for class swapping[213].

MCC is a contingency matrix method of calculating the Pearson product-moment correlation coefficient between actual and predicted values:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{FN} + \mathrm{FP})}} \quad (\mathrm{II}.41)$$

In a multi-class classification setting the Matthews correlation coefficient can be defined in terms of a confusion matrix $\mathrm{C}$ for $\mathrm{K}$ classes:

$$\mathrm{MCC} = \frac{\mathrm{c} \times \mathrm{s} - \sum_{\mathrm{k}}^{\mathrm{K}} \mathrm{p_k} \times \mathrm{t_k}}{\sqrt{(\mathrm{s}^2 - \sum_{\mathrm{k}}^{\mathrm{K}} \mathrm{p_k^2})\,(\mathrm{s}^2 - \sum_{\mathrm{k}}^{\mathrm{K}} \mathrm{t_k^2})}} \quad (\mathrm{II}.42)$$

With:

- $\mathrm{c} = \sum_{\mathrm{k}}^{\mathrm{K}} \mathrm{C_{kk}}$ : the total number of elements correctly predicted.
- $\mathrm{s} = \sum_{\mathrm{i}}^{\mathrm{K}} \sum_{\mathrm{j}}^{\mathrm{K}} \mathrm{C_{ij}}$: the total number of elements.
- $\mathrm{p_k} = \sum_{\mathrm{i}}^{\mathrm{K}} \mathrm{C_{ki}}$: the number of times that class k was predicted.
- $\mathrm{t_k} = \sum_{\mathrm{i}}^{\mathrm{K}} \mathrm{C_{ik}}$: the number of times that class k truly occurred.

MCC ranges in the interval [ −1, +1], with extreme values –1 and +1 reached in case of perfect misclassification and perfect classification, respectively, while MCC = 0 is the expected value for the coin tossing classifier.

# III. NATURAL LANGUAGE PROCESSING FOR PUBLIC HEALTH SURVEILLANCE: THE TARPON PROJECT

## III.1 TARPON: Context

### III.1.1 Project aim

The TARPON project (Traitement Automatique des Résumés de Passage aux urgences dans le but de créer un Observatoire National, aims to demonstrate the feasibility of setting up a national observatory of trauma. We propose, hereby, to compare the performances of several models for the classification of ED visits for trauma based on clinical notes from the adult emergency department of the Bordeaux University Hospital. We compared the transformers FlauBERT, CamemBERT, BelGPT2 and a French GPT2 model pre-trained on a domain-specific corpus called GPTanam here to several traditional machine learning classifiers. To the best of our knowledge, no previous performance evaluation of multiple transformers for a classification application has been conducted on complex and unstructured clinical data from ED combining common French language, medical data and jargon.

When related to clinical notes, CNN, RNN and BERT models have often shown greater performance than traditional machine learning methods [133,134]. However, when related to French ED clinical notes, Metzger et al. [62] found that traditional methods were more accurate than neural network. Models used for comparing traditional machine learning methods are often shallow neural networks, therefore we aimed at comparing Transformers models with a large panel of traditional classifiers.
The corpora on which models are trained also has an influence on NLP tasks performances [150,151,153]; we aimed at assessing the influence of the self-supervised training corpus and of a supplementary domain-specific supplementary self-supervised training step as shown by *Valmianski et al*[187] for ED clinical notes.
Furthermore, the size of Transformer models has been shown to not necessarily improve performance on NLP tasks[44], we aimed at assessing these findings on our clinical notes.

## III.2 TARPON: Methods

### III.2.1 Medical ethics regulations and GDPR

This study was authorized by the Bordeaux University Hospital Ethical Board under number GP-CE2021-21. A data management plan was created and reviewed by the privacy security board to meet institutional and national requirements in French for GDPR compliance.

### III.2.2 Database

The clinical notes were extracted from the EHR of the adult emergency department stored in the information system of the University Hospital of Bordeaux, France. They correspond to 375,478 medical records of visits to the adult emergency department of Bordeaux Hospital

from 2012 to 2020. The variables available were age, sex, date and time of the visit, the clinical note generated by the doctors/interns and the clinical note written by the triage nurses.

### III.2.3 Exploratory text analysis

Before modeling, we performed an exploratory text data analysis of the full corpus which comprised 390653 records.

#### *III.2.3.1. Distributions*

We started our investigation with an assessment of the availability of clinical notes depending on categories in the TARPON database. We used a chi square test to evaluate the hypothesis that distribution of missing values is equal among both categories.

#### *III.2.3.2. Length*

We pursued with an exploration of document length, focusing on the average number of words per category of healthcare provider (nurse or physician). The text of each clinical note was tokenized using the python Gensim package[214]. The average length of the notes was calculated. Since clinical notes are written by both a nurse and a physician for a given patient, a paired t-test was performed for the records where both clinical notes were available (n=305697). Since clinical notes provide detailed information about a patient encounter, it makes sense that the length of a note would be strongly associated with the severity and complexity of a patient's condition or treatment. To account for this variability, we performed an additional analysis by removing notes with an outlier word count.
Outliers were identified using the median absolute deviation (MAD). Outliers were selected at a threshold of ±3 MAD.

#### *III.2.3.3. Vocabulary*

In a second set, we investigated the vocabulary of each of the clinical note categories. We first identified the total number of words and the set of unique words for each category. Next, we calculated the symmetric difference between the set of unique words for each pair of categories. The symmetric difference provides the number of terms that appear in either category, but not both. This value was then normalized to the total size of the two vocabularies, providing a measure of the proportion of overlap of terms used between two categories. With this metric, a value of 1 would represent two completely distinct vocabularies, while a value of 0 would indicate that the two vocabularies are identical.

#### *III.2.3.4. Linguistic Features in clinical notes*

##### Part-of-Speech Tagging

As our prior analyses demonstrated differences between the length and the vocabulary of each category, we moved to investigate the possibility of linguistic differences between the

categories. We first analyzed the distribution of the parts of speech used for each word across the different note categories. Part-of-speech (POS) tags describe the characteristic structure of lexical terms within a sentence or text; therefore, it can be used for making assumptions about semantics. Other applications of POS tagging include Named Entity Recognition, Co-reference Resolution, Speech Recognition[215]. Identifying deviations between the part of speech distributions between the two different note categories would further support the notion that additional consideration must be given to the source of clinical data in order to provide accurate contextual analysis. We used the french-camembert-postag-model[216] which is a part of speech tagging model for French that was trained on the the French Tree Bank (FTB) dataset[217]. The FTB is a unique, richly annotated (and manually validated) lexical and syntactic resource for linguists and NLP[218] built by the 'Formal Linguistic Laboratory' (Laboratoire de Linguistique Formelle). The base tokenizer and model used for training the french-camembert-postag-model is 'camembert-base'. The comparison between note categories was performed using the paired t-test, with all low frequency tags (under an expected value of 2%) pooled into another category.

### Cosine Similarity

Next, to make the vocabulary analysis presented in the note structure section more formal, we compared the vocabularies of the nursing and medical notes using cosine similarity. The most straightforward and effective method was to use a powerful model (e.g. transformer) to encode sentences to get their embeddings and then use a similarity metric (e.g. cosine similarity) to compute their similarity score[219]. We used the distiluse-base-multilingual-cased-v1[220] available on SentenceTransformer library[3]. This model is a multilingual knowledge distilled version of multilingual Universal Sentence Encoder that supports French. We encoded both our clinical notes datasets, retrieved the embeddings and computed the cosine similarity between each dataset embeddings to measure the average semantic similarity of two categories. We broke the analysis down further, calculating the similarity between individual notes written by both health care providers for a given patient.

### Topic Modeling of clinical notes

Topic modeling was performed with the help of BERTopic[96] which is a topic model that extracts coherent topic representation through the development of a class-based variation of TF-IDF. More specifically, BERTopic generates document embedding with pre-trained transformer-based language models, clusters these embeddings, and finally, generates topic representations with the class-based TF-IDF procedure. A Uniform Manifold Approximation and Projection for Dimension Reduction[221] (UMAP) and a Hierarchical Density-Based Spatial Clustering of Applications with Noise[97] (HDBSCAN) are used to, respectively, reduce the dimensionality of document embeddings and model clusters.

[3] https://www.sbert.net/index.html

### III.2.4 Labeling strategy

69,110 clinical notes were randomly extracted for manual annotation. Our coding team consisted of trauma epidemiologists, emergency physicians, emergency nurses, research assistants, and biostatisticians, for a total of 16 coders. The annotation phase lasted 5 months. For each clinical note, a code describing the content of the text was assigned. The annotation grid Appendix J-K used for the coding was developed for the needs of the project. The code associated with each clinical notes consisted of 9 fields. The fields were: "First visit (to the emergency department for this reason)", "Location (of the trauma)", "Activity (performed during the trauma)", "Type of Sport (practiced during the trauma)", "Subject under the influence", "Notion of pre-traumatic discomfort", "MVA (Motor Vehicle Accident)-Secondary Prevention Elements", "MVA-Antagonist", "Type of trauma or Mode of travel for the MVA". The objective being to classify the types of trauma, we used mainly the data of the field "Type of trauma or Mode of movement for the MVA". The distribution of the latter being unbalanced, we created a composite variable containing 8 mutually exclusive classes in order to have a larger number of clinical notes per class. Therefore, we grouped certain types of trauma (i.e. "Fall" which included "Fall from own height," "Fall from a given height," and "Fall on stairs"). The composite variable included the following classes/labels: "Accident of exposure to body fluids (blood exposure accident, unprotected sex at risk)" (AEF), "Assault", "Motor Vehicle Accident (MVA)", "Foreign body in eyes" (FBE), "Fall (except sports)", "Sports accident" (Acc. sport), "Intentional Injury", "Other trauma" as seen in Figure III.1. The inter-annotator agreement was assessed with a random sample of 1000 clinical notes labelled by two annotators leading to a Cohen's kappa score[222] of 0.84.

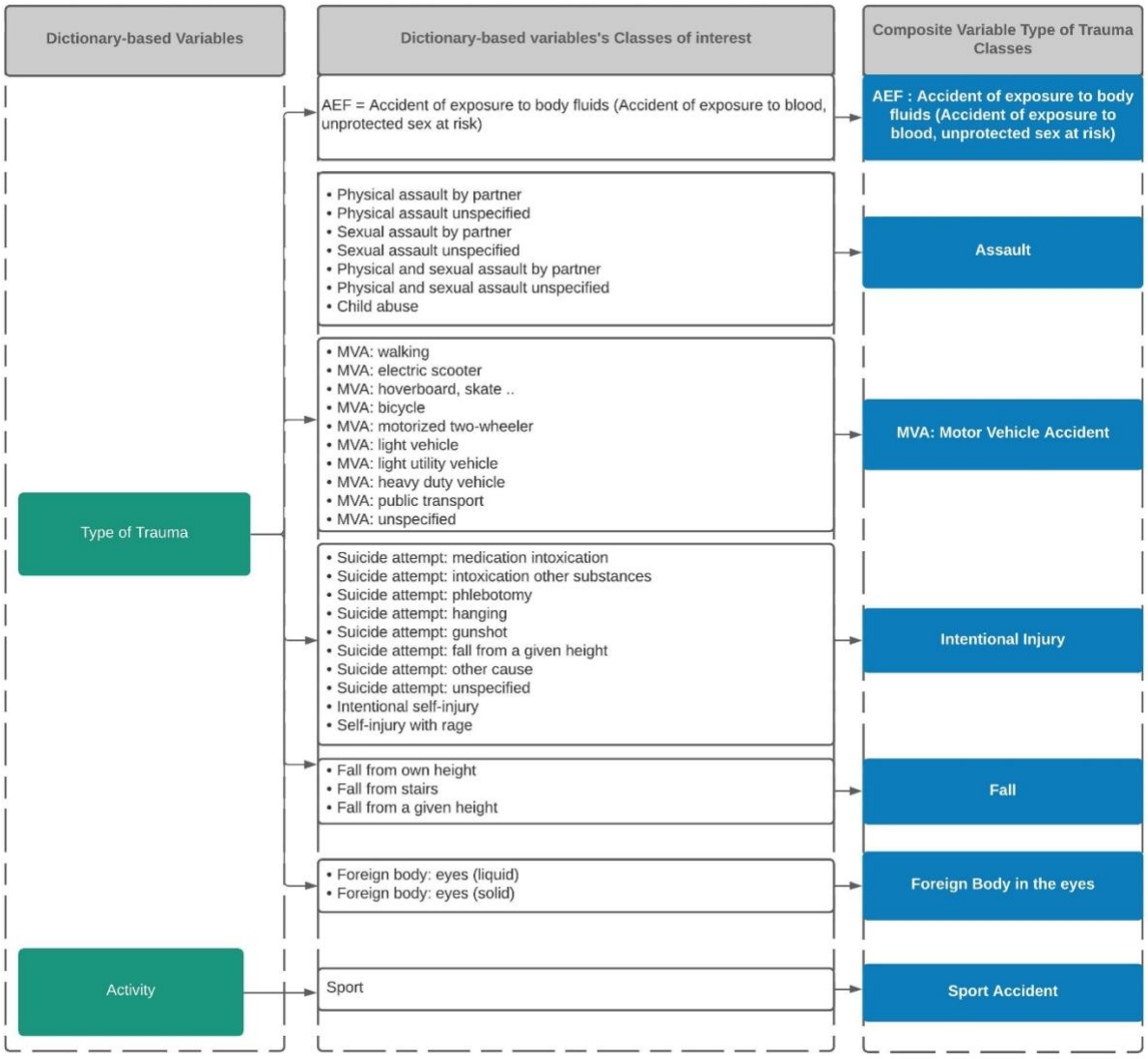



*Figure III.1 Composite variable type of trauma based on the annotation grid variables.*

In total, 22,481 manually labeled clinical notes from Bordeaux University Hospital were included in the study. Indeed, one-third (22,481/69,110) of the total annotated clinical notes were labeled as visit to the ED resulting from a trauma. The average number of sentences of the corpus was of 3.25 (min:1, max:63, std:2.56). The average length of clinical notes was of 58 words with a minimum of 1 (e.g., AES, Accident d'exposition au sang), a maximum of 630 and a standard deviation of 38 words. Unique unigrams, bigrams and trigrams were respectively equal to 70499, 395827 and 777459.

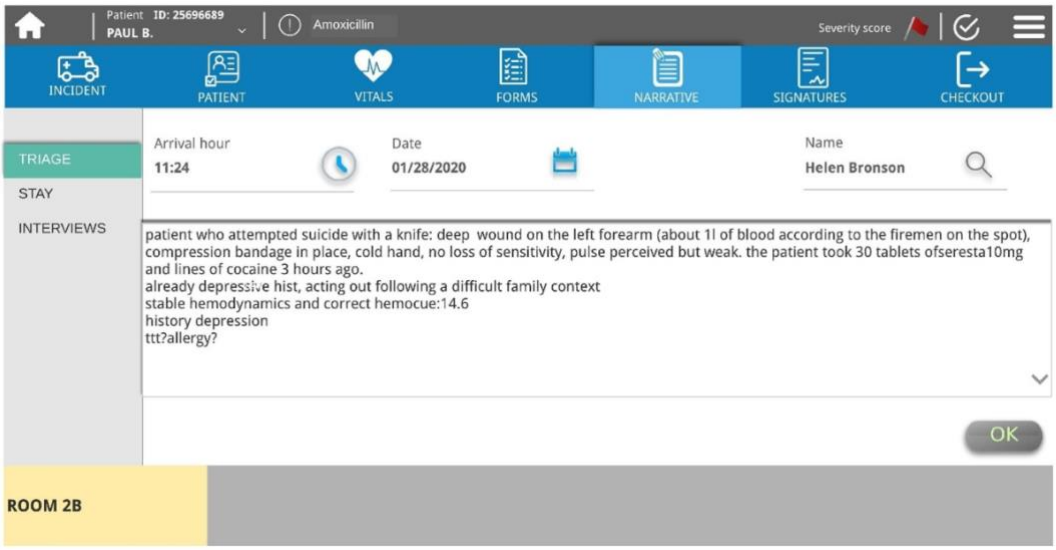



*Figure III.2 Example of a clinical note*

A sensitivity analysis was performed in order to study the impact of potentially ambiguous content as regard to its classification. Therefore, the test sample was re-read by an expert. Potentially ambiguous content as regard to its classification is defined here as the accumulation of several mechanisms or types of trauma and/or a major difficulty in assigning a label to a clinical note given its text.

### III.2.5 Models and experiment settings

The models selected for comparison and freely available as open-source content were:

- TF-IDF with several classifiers
- Transformers with different (further details in section III.2.5.3):
    - Architecture: BERT and GPT-2
    - Sizes: number of heads and dimensions
    - Pre-training corpora languages and tokenizers: OSCAR, Wiki and CCNET
    - Pre-training step on domain specific corpora strategy

#### *III.2.5.1. Pre-processing*

All clinical notes were:

- lower-cased
- punctuation stripped, since clinical texts often exhibit varying levels of fragmentation and grammatical correctness (gensim)
- non alphanumerical characters stripped (gensim)
- multiple whitespaces stripped (gensim)
- stop-words removed with the French nltk[223] stopwords corpus to which we added the single letters 'h' (abbreviated term for hour), 'g' (for grams) and 'a'.

The proper handling of numerical elements in text remains an open question in the NLP community. This transformation has the potential to incorrectly bias the interpretation of analyses, particularly those which use vocabulary similarities or rely on normalized word

frequencies. We did not strip the digits for the next step of the work since we plan to use semantic representations.

### Vectorization and feature extraction for traditional machine learning

To tokenize, count occurrences and normalize the raw text of each clinical note, the sci-kit learn class TfidfVectorizer[224] was used. The encoding was left as default to "utf-8". The scikit learn formula for TF-IDF is as follow:

$$\mathrm{idf}(t) = \log\frac{1+n}{1+\mathrm{df}(t)} + 1 \qquad (III.1)$$

where $n$ is the total number of documents in the document set, and $\mathrm{df}(t)$ is the number of documents in the document set that contain term t.
The effect of adding "1" to the idf in the equation above is that terms with zero idf, i.e., terms that occur in all documents in a training set, will not be entirely ignored.
The resulting tf-idf vectors are then normalized by the Euclidean norm:

$$v_{norm} = \frac{v}{||v||_2} = \frac{v}{\sqrt{{v_1}^2 + {v_2}^2 + \ldots + {v_n}^2}} \qquad (III.2)$$

#### *III.2.5.2. Traditional machine learning classifiers*

As seen in section II.2.3.6, SVM has been shown to have the best performance for a classification task on EHR clinical notes, however, Metzger et al. found that SVM might not be the best classifier for French ED clinical notes. Therefore, and as we wanted our analysis to be exhaustive, we tested 17 different machine learning classifiers listed below. For each classifier, the best parameters of estimators were searched exhaustively using GridSearchCV[4] class from scikit learn. Most of the models were provided by scikit-learn library. The time for execution (in seconds) was calculated for each model.

### Linear models tested were:

- Logistic Regression method despite its name, is a linear model for classification rather than regression[225]. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function. Here, logistic regression was used with the 'elasticnet' penalty which uses both l1 and l2 penalties, the l1 ratio was set to 0.5. The solver chosen was 'saga', tolerance was left as default to $1e^{-4}$
  Since the classes were imbalanced, the 'weight_class' parameter was set to 'balanced' which uses the values of $y$ to automatically adjust weights inversely proportional to class frequencies in the input data as $n_{samples}/(n_{classes} * \mathrm{np.bincount}(y))$

[4] https://scikit-learn.org/stable/modules/grid_search.html

- Linear Support Vector Machines for Classification (SVC) which were described in section II.2.3.4. The aim is a risk minimization for the equation with a “one-versus-one” approach for multi-class classification[226,227]:

$$C\sum_{i=1,n}\mathcal{L}(f(x_i),y_i)+\Omega(w) \qquad (III.3)$$

where:
- $x_i \in R^p$ are training vectors.
- $w \in R^p$ are weights.
- $C$ is used to set the amount of regularization.
- $\mathcal{L}$ is a loss function of our samples and the model parameters.
- $\Omega$ is a penalty function of the model parameters.

The penalty used was l2 and the loss was hinge where the following primal problem is solved:

$$\min_w \frac{1}{2}w^T w + C\sum_{i=1}^{1}\left(\max(0,1-y_i w^T x_i)\right)^2 \qquad (III.4)$$

LinearSVC implements “one-vs-the-rest” multi-class strategy. Tolerance was set to $1e^{-5}$ and the class weight was set to 'balanced'.

- Stochastic Gradient Descent (SGD) which is a simple yet very efficient approach to fitting linear classifiers under convex loss functions such as (linear) Support Vector Machines and Logistic Regression[228,229]. Stochastic gradient descent is an optimization method for unconstrained optimization problems. In contrast to (batch) gradient descent, SGD approximates the true gradient of $E(w,b)$ by considering a single training example at a time. The class SGDClassifier implements a first-order SGD learning routine. The algorithm iterates over the training examples and for each example updates the model parameters according to the update rule given by:

$$w \leftarrow w - \eta\left[\alpha\frac{\partial R(w)}{\partial w}+\frac{\partial L(w^T x_i+b,y_i)}{\partial w}\right] \qquad (III.5)$$

where $\eta$ is the learning rate which controls the step-size in the parameter space. The intercept $b$ is updated similarly but without regularization.
On our dataset, after an hyperparameters search, it has been used with its default loss function which is 'hinge', returning a linear SVM (Support Vector Machines), the penalty was 'l2' and class_weight was set to 'balanced'.

- Perceptron[230] is a simple classification algorithm suitable for large scale learning. By default, it does not require a learning rate, it is not regularized (penalized), it updates its model only on mistakes. The parameters were as follow: alpha to 0.0001, no penalty, class_weight to 'balanced', tolerance to '0.001', number of iterations with no improvement to wait before early stopping to 10.

- Passive Aggressive classifier[231] is an online learning algorithm. Each instance is represented by a vector and the prediction mechanism is based on a hyperplane which divides the instance space into two half-spaces. The margin of an example is proportional to the distance between the instance and the hyperplane. The PA algorithm utilizes the margin to modify the current classifier. The update of the classifier is performed by solving a constrained optimization problem. The loss function used was hinge which is equivalent to PA-I in the paper.

- Ridge Classifier with Cross Validation[232] first converts binary targets to $-1,1$ and then treats the problem as a regression task, optimizing the same objective as above. The predicted class corresponds to the sign of the regressor's prediction. For multi-class classification, the problem is treated as multi-output regression, and the predicted class corresponds to the output with the highest value. It might seem questionable to use a (penalized) Least Squares loss to fit a classification model instead of the more traditional logistic or hinge losses. However, in practice, all those models can lead to similar cross-validation scores in terms of accuracy or precision/recall, while the penalized least squares loss used by the RidgeClassifier allows for a very different choice of the numerical solvers with distinct computational performance profiles. The RidgeClassifier can be significantly faster than e.g. Logistic Regression with a high number of classes because it can compute the projection matrix $(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$ only once. This classifier is sometimes referred to as a Least Squares Support Vector Machines with a linear kernel. The parameters were set to 'sparce_cg' as solver, a 5-fold cross-validation and class weights to 'balanced'.

### The leveraged trees classifiers were:

- Decision Trees classifier[119] is a non-parametric supervised learning method. Under the method "DecisionTreeClassifier", scikit-learn uses an optimized version of the CART (Classification And Regression Tree) algorithm as described in section II.2.3.3. This algorithm was used with a 'gini' loss criterion and a max depth of 14 nodes.

- Extra Trees algorithm[233] builds an ensemble of unpruned decision or regression trees according to the classical top-down procedure. Its two main differences with other tree-based ensemble methods are that it splits nodes by choosing cut-points fully at random and that it uses the whole learning sample (rather than a bootstrap replica) to grow the trees. The Extra-Trees algorithm builds an ensemble of unpruned decision or regression trees according to the classical top-down procedure. Its two main differences with other tree-based ensemble methods are that it splits nodes by choosing cut-points fully at random and that it uses the whole learning sample (rather than a bootstrap replica) to grow the trees.

### Nearest Neighbors tested was:

- K-Nearest Neighbors classifier has been described in section II.2.3.2. It has been used with 50 neighbors and the Euclidean distance as weights.

Naive Bayes algorithm assessed was:

- Complement Naive Bayes classifier as available with scikit-learn was proposed by Rennie and al. in 2003[234]. This algorithm corrects two specific problems. First of all, when one class has more training examples than another, Naive Bayes selects poor weights for the decision boundary. Secondly, with Naive Bayes, features are assumed to be independent. As a result, even when words are dependent, each word contributes evidence individually. Thus, the magnitude of the weights for classes with strong word dependencies is larger than for classes with weak word dependencies. When applied to text data, transforming multinomial naïve bayes on term and document frequencies in addition to a transformation based on length of the document improves the performance of the regular naïve bayes classification algorithm. For our dataset, alpha was left to 0, force_alpha was left to True, and norm was set to True.

We also explored ensemble methods such as:

- Random Forest classifier[119] is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. For our dataset, after a GridSearch, the criterion chosen was 'gini', the maximum depth of the trees was set to None.

- Bagging algorithm as defined in section II.2.3.5 was used with the two most performing classifiers: the Linear Support Vector classifier and the Stochastic Gradient Descent classifier. We kept the same parameters for both classifiers as used in previous settings without bagging. The number of estimators was set to 500.

- AdaBoost classifier[123] is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. The sci-kit learn library's adaboost classifier implements the algorithm known as AdaBoost-SAMME. After hyper-parameters tuning (245 minutes), the settings for the Decision-Tree classifier (base estimator) were 10 for the maximum depth of the tree and 5 for the minimum number of samples required to be at a leaf node. The AdaBoost had a learning rate of 0.01 and 10 estimators.

- Histogram Gradient Boosting classifier is a histogram-based Gradient Boosting Classification Tree provided by the scikit-learn library and is inspired by LightGBM[235]. These histogram-based estimators can be orders of magnitude faster than the classic Gradient Boosting Classifier. These fast estimators first bin the input samples $X$ into integer-valued bins (typically 256 bins) which tremendously reduces the number of splitting points to consider and allows the algorithm to leverage integer-based data structures (histograms) instead of relying on sorted continuous values when building the trees.

- Extreme Gradient Boosting classifier[236] provides a parallel tree boosting. This algorithm is available with the xgboost library[5]. After hyper-parameters tuning the algorithm had the following settings: the maximum depth was of 9, the minimum children weight (minimum sum of instance weight (hessian) needed in a child) was of 1, gamma (minimum loss reduction required to make a further partition on a leaf node of the tree) was set to 0.1, the subsample (denotes the fraction of observations to be randomly samples for each tree) was set to 0.8, the columns sample by tree was set to 0.8, alpha (L1 regularization term on weight) was set to 1.

- Light Gradient Boosting Machine[237] was introduced in 2017. LightGBM uses histogram-based algorithms, which bucket continuous feature (attribute) values into discrete bins. This speeds up training and reduces memory usage. LightGBM has many of Boost's advantages, including sparse optimization, parallel training, multiple loss functions, regularization, bagging, and early stopping. A major difference between the two lies in the construction of trees. LightGBM does not grow a tree level-wise — row by row — as most other implementations do. Instead, it grows trees leaf-wise. It chooses the leaf it believes will yield the largest decrease in loss. Besides, LightGBM does not use the widely used sorted-based decision tree learning algorithm, which searches the best split point on sorted feature values, as Boost or other implementations do. Instead, LightGBM implements a highly optimized histogram-based decision tree learning algorithm, which yields great advantages on both efficiency and memory consumption. The LightGBM algorithm utilizes two novel techniques called Gradient-Based One-Side Sampling and Exclusive Feature Bundling which allow the algorithm to run faster while maintaining a high level of accuracy. This algorithm has been used with a learning rate of 0.1, a max number of leaves in one tree of 30, a maximum bin (max number of bins that feature values will be bucketed in) of 255 on 100 iterations.

#### III.2.5.3. Deep Learning algorithms

Transformers models pre-trained on French corpora were chosen. In order to test the effect of different corpora we used the CamemBERT[69] pre-trained with either OSCAR[238], Wikipedia or CCNET.
In order to test the effect of the model sizes, we used FlauBERT small, cased and large. The main results are presented for the most performing model among CamemBERT and FlauBERT models.
The GPT2 model which was trained on a French corpus and named BeglGPT2 was also used.
We then chose the most performing model (BelGPT2) and fine-tuned it with the remaining 306,368 unlabeled clinical notes, this model being called, here, GPTanam.
For all transformers the optimizer was AdamW with an epsilon of 1e-8 and the maximum length was of 512. GPTanam had training and evaluation batch sizes of 5 and the learning rate was of $2e^{-5}$. For FlauBERT and CamemBERT, batch sizes for training was of 16 and 20 for

[5] https://xgboost.readthedocs.io/en/latest/python/index.html

evaluation and the learning rate was of 5e-5. Models were trained with the hugging face library under Pytorch framework on our workstation with a single Titan RTX (Nvidia©) GPU with 24GB of VRAM. Performance analysis was performed with scikit-learn and imbalance-learn v0.9.1

### III.2.6 Self-supervised learning and Fine-tuning phase

Considering the GPTanam model, a first step comprising a self-supervised learning was performed with 306,368 clinical notes with one epoch[239]. For all models, a random sample of 80% (n= 18166) of the labeled as trauma (n=22481) was dedicated to supervised learning. This dataset was divided into a training sample (n=14532) and a validation sample (n=3634) with an 80/20 ratio. We trained each model 9 times with different seeds on 7 epochs for CamemBERT and FlauBERT models and 5 epochs for BelGPT2 and GPTanam. In order to obtain a single prediction for the 9 different executions of the chosen epoch (based on maximum validation micro F1-score) for each model, a vote was taken.

### III.2.7 Test phase

The test sample contained 20% of the labeled dataset, i.e. 4315 records. The second reading of these clinical notes resulted in 467 being tagged as clinical notes with potentially complex and/or ambiguous content as regard to its classification. The analysis therefore included both the complete test dataset (n=4315) and the dataset without complex and/or ambiguous content (n=3848). In order to obtain the probabilities for each prediction, a softmax activation layer was applied to the 4 transformer models.

### III.2.8 Labeled datasets

The label distribution among the corpus and each train, validation and test dataset is presented in Table III.1. The most common type of trauma was the class "Fall" followed by "Other trauma" and "Motor Vehicle Accident". An example of clinical notes translated from French is given Figure III.3.

| | |
|---|---|
| Patient emmené par les pompiers pour chute à domicile, TC sans PC. Sous kardegic. Chute mécanique, lésion du scalp. | Patient brought by the fire department for a fall at home with CT without LOC. Under kardegic. Mechanical fall scalp lesion. |
| AVP VL/VL, impact frontal, faible vitesse, airbag déclenché, pas de PC, pas PCI, pas dlr cou, pas céphalées | MVA LV/LV, frontal impact, low speed, airbag activated, no CT, no ILOC, no neck pain, no headache |

*Figure III.3 Example of clinical notes*

| Type of Trauma | Train dataset | | Validation dataset | | Test dataset | | Total | |
|---|---|---|---|---|---|---|---|---|
| | n | % | n | % | n | % | n | % |
| Accident of Exposure to Bodily Fluids | 132 | (0.9%) | 40 | (1.1%) | 41 | (1%) | 213 | (0.9%) |
| Assault | 1587 | (10.9%) | 393 | (10.8%) | 498 | (11.5%) | 2478 | (11%) |
| Fall | 4778 | (32.9%) | 1162 | (32%) | 1554 | (36%) | 7494 | (33.3%) |
| Foreign Body in Eye | 642 | (4.4%) | 180 | (5%) | 186 | (4.3%) | 1008 | (4.5%) |
| Intentional Injury | 341 | (2.3%) | 73 | (2%) | 112 | (2.6%) | 526 | (2.3%) |
| MVA | 2028 | (14%) | 495 | (13.6%) | 568 | (13.2%) | 3091 | (13.7%) |
| Other trauma | 3713 | (25.6%) | 950 | (26.1%) | 985 | (22.8%) | 5648 | (25.1%) |
| Sport Accident | 1311 | (9%) | 341 | (9.4%) | 371 | (8.6%) | 2023 | (9%) |
| Total | 14532 | (64.6%) | 3634 | (16.2%) | 4315 | (19.2%) | 22481 | (100%) |

*Table III.1* Labels Distribution among Train, validation and test dataset. MVA: Motor Vehicle Accident

On the labeled dataset, the median age at the visit was 37 years (1st and 3rd quartiles [24–58]) and 58.5% of patients were male. Electronic health record was introduced in year 2012 in Bordeaux University hospital, which explains the lower proportion of data for this particular year. Year 2019 saw a decrease in ED venues while in 2020 there have been a significant increase. Table 3 summarizes the characteristics of the train, validation and test datasets for the concerned population. Distribution of the variables age, sex and year of venues at the ED were comparable among the 3 datasets.

| | **Train dataset** | | **Validation dataset** | | **Test** | | **Total** | |
|---|---|---|---|---|---|---|---|---|
| | **n** | **%** | **n** | **%** | **n** | **%** | **n** | **%** |
| **Age** | 37 | (24-58) | 37 | (24-57) | 37 | (24-58) | 37 | (24-58) |
| **Sex. Male** | 8486 | (58.3%) | 2181 | (59.9%) | 2476 | (57.5%) | 13143 | (58.5%) |
| **Year of ED venue** | | | | | | | | |
| **2012** | 218 | (1.9%) | 52 | (1.8%) | 66 | (1.9%) | 336 | (1.9%) |
| **2013** | 1389 | (12.2%) | 359 | (12.4%) | 418 | (12.3%) | 2166 | (12.2%) |
| **2014** | 1444 | (12.6%) | 385 | (13.3%) | 386 | (11.3%) | 2215 | (12.3%) |
| **2015** | 1502 | (13.1%) | 326 | (11.2%) | 425 | (12.5%) | 2253 | (12.6%) |
| **2016** | 1419 | (12.4%) | 365 | (12.6%) | 426 | (12.6%) | 2210 | (12.3%) |
| **2017** | 1493 | (13.1%) | 370 | (12.8%) | 461 | (13.5%) | 2324 | (12.9%) |
| **2018** | 1425 | (12.5%) | 405 | (13.9%) | 474 | (13.9%) | 2304 | (13.5%) |
| **2019** | 690 | (6%) | 175 | (6%) | 218 | (6.4%) | 1083 | (6.2%) |
| **2020** | 1856 | (16.2%) | 468 | (16.1%) | 532 | (15.6%) | 2856 | (16%) |
| **Missing values** | 3118 | (27.3%) | 737 | (25.4%) | 899 | (26.4%) | 4724 | (20.9%) |

*Table III.2* Train, validation and test dataset characteristics

Numbers are given along with percentages for sex and year of emergency department venue variables. Median, first quartile and fourth quartile are given for age.

### III.2.9 Error analysis

An error analysis was performed with uni and bigrams for the best performing model. All clinical notes misclassified were read by an expert to determine whether the human annotation label was appropriate or not.

## III.3 TARPON: Results

### III.3.1 Clinical notes' structure

#### *III.3.1.1. Missing value distribution*

Firstly, the Table III.3 shows the distribution of missing clinical notes among categories and at the intersection of both categories for a given patient. Percentages are expressed in parentheses. There were statistically (CI:95) more missing notes for emergency physicians than for nurses (p<0.001).

| Clinical notes | Triage Nurses | Emergency Physician | Both Nurse and Physician |
|---|---|---|---|
| Available | 373728 (99.53%) | 295570 (78.72%) | 2938020 (78.25%) |
| Missing | 1750 (0.47%) | 79908 (21.28%) | 81658 (21.75%) |

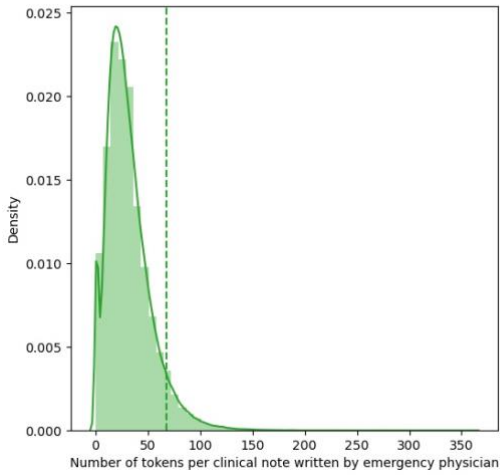
*Table III.3* Availability of clinical notes in the TARPON database

#### *III.3.1.2. Length*

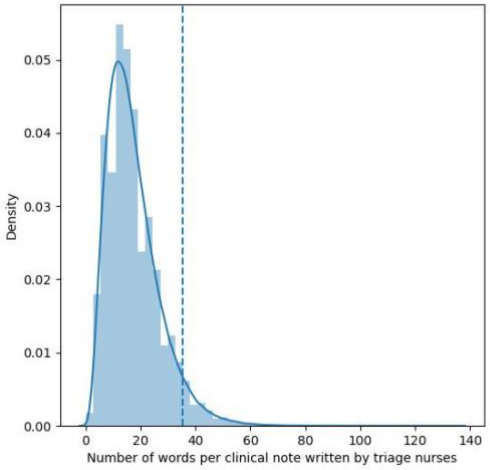
The results for narratives length evaluation can be found in Table III.4. Pairs of nursing and medical clinical notes had statistically significant different lengths at 95% confidence for both the complete set of notes and for the set excluding notes with an outlier number of tokens. Distributions of the number of tokens per clinical notes can been seen on Fig XX.

| Note type | Mean word count | Word Count SD | Total Notes |
|---|---|---|---|
| Triage Nurses | 17.34 (15.97) | 6.61 (7.55) | 373728 (354955) |
| Emergency Physicians | 31.47 (27.75) | 21.60 (15.67) | 295570 (293820) |

*Table III.4* Average Document Length for both the complete set of notes and for notes excluding those with an outlier number of tokens (in parentheses)




(a) Distribution of the Number of Tokens per clinical notes written by emergency physicians




(b) Distribution of the Number of Tokens per clinical notes written by triage nurses

*Figure III.4 Distribution of the Number of Tokens per clinical notes categories.*

*The dashed vertical lines represent the +3 Mean Absolute Deviation threshold.*

#### *III.3.1.3. Vocabulary*

Next, Table III.5 provides the results for the vocabulary analysis. Within this table the first two columns present the normalized symmetric differences between the unique tokens found in each category. The symmetric difference is the set of terms existing in either category, but not both, and is normalized by the total unique word count of both categories. The final two columns represent the count of total words, and unique words present in each category respectively. The detail analysis of the words not appearing in both sets reveals that those words are mostly misspelled or that spaces between words are missing (i.e. "abcesde', 'abcesinflammatoire', 'abcesl', 'abcse'). Cosine similarity (as defined in section III.2.3.4) between the 2 corpus was of 0.87, and the Jaccard distance was of 0.72.

| | **Normalized Symmetric Difference (%)** | | | |
|---|---|---|---|---|
| | Triage Nurses | Emergency Physicians | Total Words | Total Unique Words |
| Triage Nurses | 0 | 0.56 | 6,741,895 | 111,765 |
| Emergency Physicians | 0.56 | 0 | 9,679,841 | 132,452 |

*Table III.5* Vocabulary differences by category

### III.3.2 Linguistic features

We started the linguistic analysis with an assessment of the differences between each categories' parts of speech distribution performed by the french-camembert-postag transfomer[240] that was trained on the free-french-treebank dataset[241] for which the matching

tags and categories are available in Table III.6. The normalized tag proportions (over 2%) are represented on Figure III.5 and Figure III.6 depicts the whole tags distribution.

| Tag | Category |
|---|---|
| ADJ | adjectif |
| ADJWH | adjectif |
| ADV | adverbe |
| ADVWH | adverbe |
| CC | conjonction de coordination |
| CLO | pronom |
| CLR | pronom |
| CLS | pronom |
| CS | conjonction de subordination |
| DET | déterminant |
| DETWH | déterminant |
| ET | mot étranger |
| I | interjection |
| NC | nom commun |
| NPP | nom propre |
| P | préposition |
| P+D | préposition + déterminant |
| PONCT | signe de ponctuation |
| PREF | préfixe |
| PRO | autres pronoms |
| PROREL | autres pronoms |
| PROWH | autres pronoms |
| U | ? |
| V | verbe |
| VIMP | verbe imperatif |
| VINF | verbe infinitif |
| VPP | participe passé |
| VPR | participe présent |
| VS | subjonctif |

*Table III.6* Parts-of-speech matching for each tag of the French Treebank dataset[218]

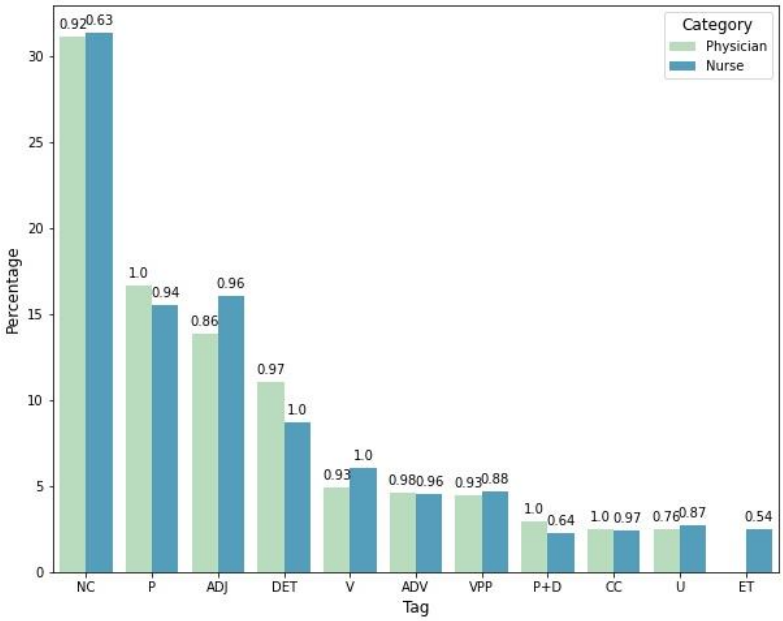



*Figure III.5 Distribution of the major Part-Of-Speech tags (over 2%) normalized on length among clinical notes for both physicians and nurses' categories (french-camembert-postag-model's confidence scores are given upon each bar)*

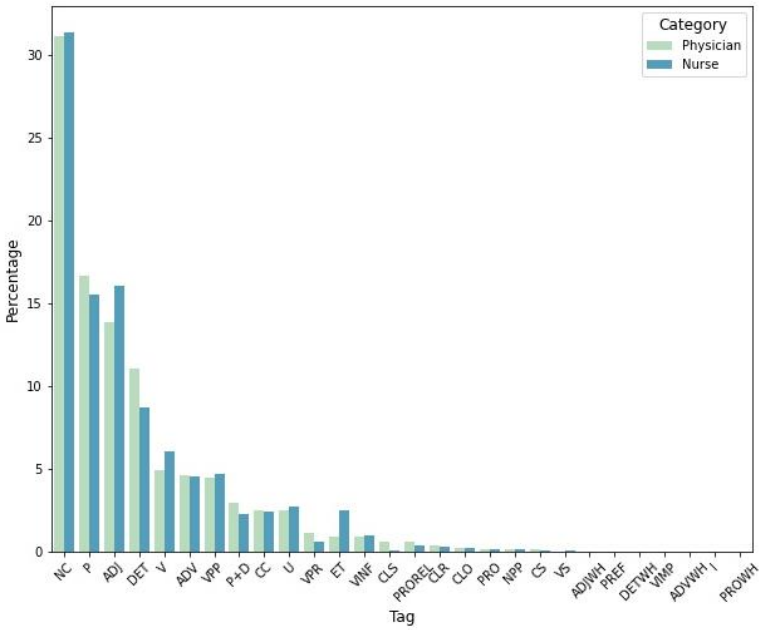



*Figure III.6 : Distribution of all the Part-Of-Speech tags normalized on length among clinical notes for both physicians and nurses' categories.*

Nurses' notes contained significantly more adjectives, verbs and abbreviations (tagged as 'U' and/or 'ET' here) but also fewer specifiers (fr: déterminants, e.g. un, une, des), prefix (fr: prépositions, e.g. dans, chez, sous).

### III.3.3 Topic Modeling

The top 5 topics as seen on Figure III.7 were:

- Symptom-based with neck pain, associated with seatbelt which could be related to Motor Vehicle Accident (MVA), broken femur or hematoma (on head)
- Based on trauma mechanism with scooter accident or assault with head trauma

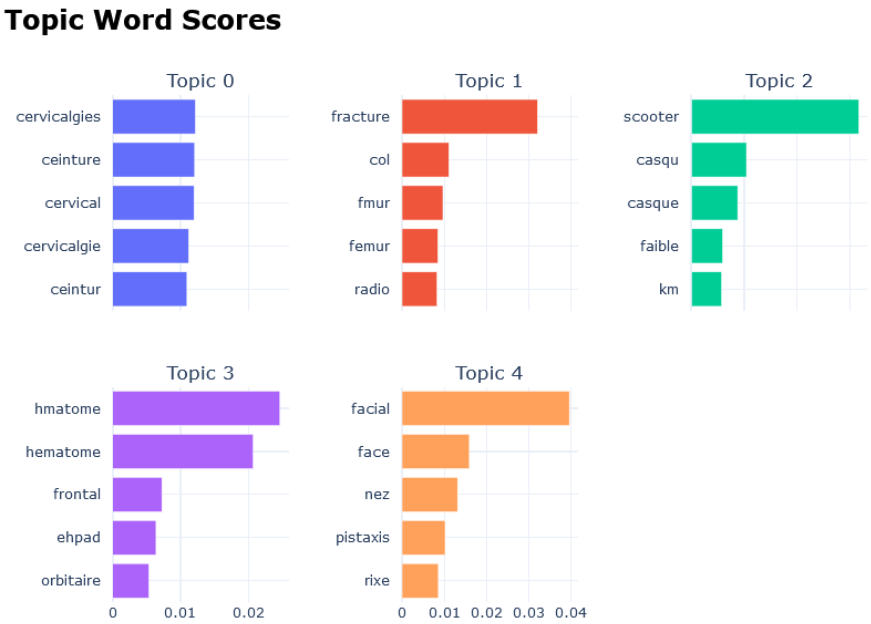



*Figure III.7 Top 5 topics identified by BERTopic with their most frequent words and scores*

When selecting 50 topics, the aggregation was based on symptoms, trauma mechanism, body parts or type of patients or sport.

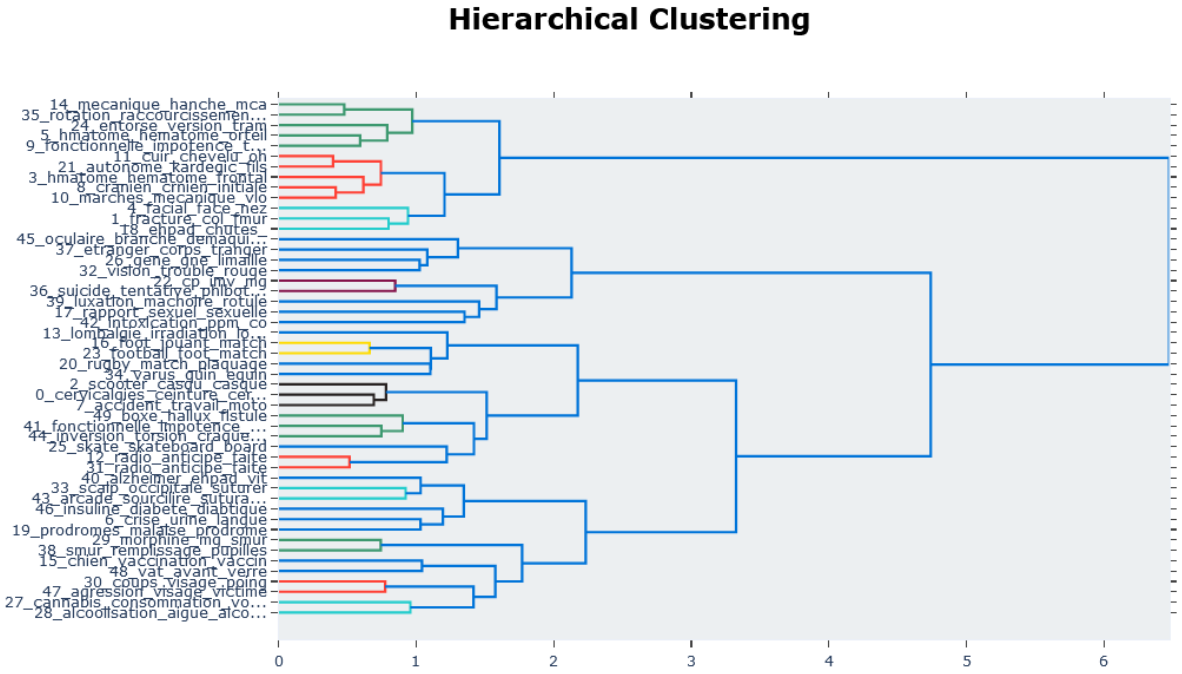



*Figure III.8 Hierarchical clustering of the top 50 topics identified by BERTopic*

### III.3.4 Fine-tuning performance of models

Unlike machine learning methods such as TF-IDF, supervised fine-tuning of the transformer models is time-consuming and is greatly accelerated with the use of GPUs (Graphics Processing Unit). The self-supervised fine-tuning step for GPTanam model took approximately 12 hours. At that point, GPTanam could generate artificial clinical notes as seen in Figure III.10, which could not easily be differentiated from the original ones.

| | |
|---|---|
| Patient emmené par les pompiers pour chute à domicile, TC sans PC. Sous kardegic. Chute mécanique, lésion du scalp. | Patient brought by the fire department for a fall at home with CT without LOC. On Kardegic. Mechanical fall, scalp lesion. |
| AVP VL/VL, impact frontal, faible vitesse, airbag déclenché, pas de PC, pas PCI, pas dlr cou, pas céphalées | MVA LV/LV, frontal impact, low speed, airbag activated, no CT, no ILOC, no neck pain, no headache. |

*Figure III.9 Example of clinical notes generated by GPTanam after self -supervised pre-training step. CT: Cranial Trauma, LOC: Loss Of Consciousness, MVA: Motor Vehicle Accident, LV: Light Vehicle*

One epoch of supervised fine-tuning took 15, 16, 15, 23, 19 and 18 minutes for, respectively, CamemBERT (all 3 pretraining corpora), FlauBERT-base, FlauBERT-small, FlauBERT-large, BelGPT2 and GPTanam. When looking deeper into each transformer model's F1-scores on the validation dataset, the Figure III.10 shows that CamemBERT reached its maximum F1-score (0.873) at Epoch 6, FlauBERT-small achieved 0.874 at epoch 5, BelGPT2 was at its peak (0.890) faster at Epoch 3 and GPTanam reached 0.980 at epoch 2. Moreover, GPTanam's F1-score on the validation dataset was the highest among the 4 transformers models. We conjecture that self-supervised step on domain-specific corpus for GPTanam contributed to a learning of the semantic representations which resulted in a faster convergence in the learning of the classification task.

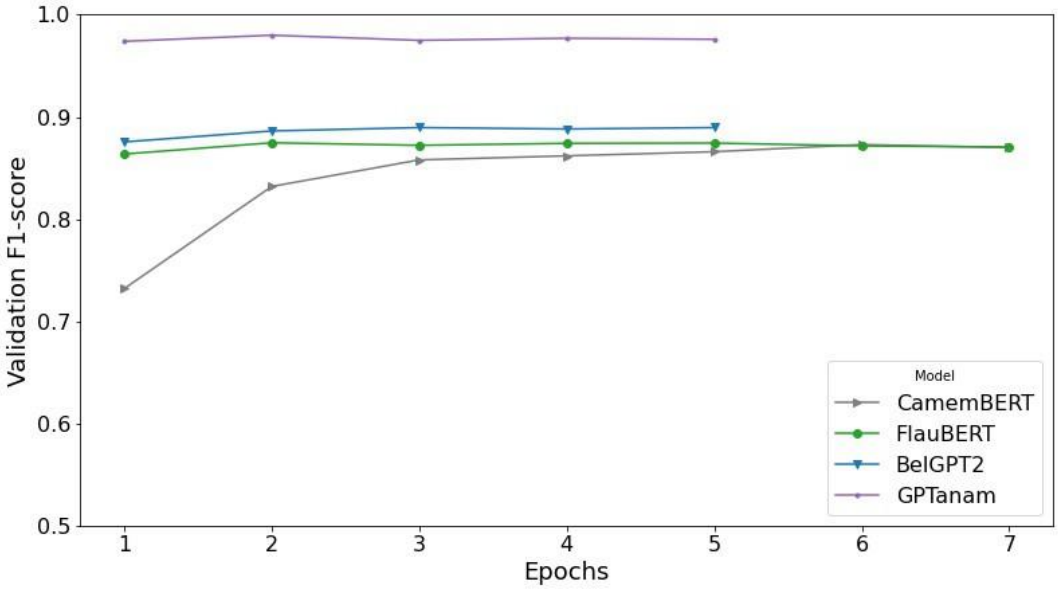



*Figure III.10* F1-score curves for CamemBERT-CCNET, FlauBERT-small, BelGPT2 and GPTanam on the validation dataset.

### III.3.5 Performance of models

When considering all models as seen on Table III.7, bagging algorithms and Light Gradient Boosting had similar results to Transformers except for GPTanam. As for CamemBERT, the pre-training corpus for transformers had a slight influence on the average micro f1-scores with a maximum gain of 0.01. A larger transformer doesn't imply better performances, and when the larger model of FlauBERT is trained for our classification task, the micro f1-score does not increase.

The following results will be kept in line with the journal article we published in "Deep Learning Transformer Models for Building a Comprehensive and Real-time Trauma Observatory: Development and Validation Study"[242], hence, the TF-IDF (Terms Frequency-Inverse Document Frequency)/SVM (Support Vector Machine) was kept as a baseline model and the best CamemBERT and FlauBERT models were kept (CamemBERT-CCNET and FlauBERT-small).

| Model | AE | Assault | Fall | Foreign body in the eye | Intent. Injury | MVA | Other | Sport accident | Average | type of classifier |
|---|---|---|---|---|---|---|---|---|---|---|
| **AdaBoost** | 0.69 | 0.83 | 0.86 | 0.29 | 0.67 | 0.89 | 0.68 | 0.33 | 0.740 | Ensemble |
| **Bagging LinearSVC** | 0.90 | 0.91 | 0.91 | 0.83 | 0.80 | 0.91 | 0.81 | 0.82 | 0.873 | Ensemble |
| **BelGPT2** | 0.83 | 0.91 | 0.92 | 0.82 | 0.77 | 0.91 | 0.85 | 0.85 | 0.887 | Transformer |
| **Bagging SGD** | 0.88 | 0.91 | 0.91 | 0.82 | 0.80 | 0.91 | 0.82 | 0.82 | 0.875 | Ensemble |
| **CamemBERT CCNET** | 0.84 | 0.91 | 0.92 | 0.84 | 0.76 | 0.90 | 0.83 | 0.83 | 0.878 | Transformer |
| **CamemBERT OSCAR** | 0.83 | 0.91 | 0.91 | 0.83 | 0.73 | 0.91 | 0.83 | 0.82 | 0.875 | Transformer |
| **CamemBERT Wiki** | 0.82 | 0.90 | 0.91 | 0.82 | 0.72 | 0.90 | 0.82 | 0.81 | 0.869 | Transformer |
| **Complement NB** | 0.90 | 0.80 | 0.86 | 0.77 | 0.74 | 0.91 | 0.70 | 0.8 | 0.816 | Naive Bayes |
| **Decision Tree** | 0.80 | 0.84 | 0.85 | 0.77 | 0.68 | 0.91 | 0.71 | 0.71 | 0.799 | Tree |
| **Extra Trees** | 0.87 | 0.90 | 0.88 | 0.83 | 0.62 | 0.92 | 0.79 | 0.65 | 0.834 | Tree |
| **FlauBERT cased** | 0.84 | 0.92 | 0.91 | 0.82 | 0.73 | 0.91 | 0.82 | 0.83 | 0.873 | Transformer |
| **FlauBERT large** | 0.86 | 0.92 | 0.92 | 0.82 | 0.71 | 0.91 | 0.82 | 0.82 | 0.876 | Transformer |
| **FlauBERT small** | 0.79 | 0.91 | 0.92 | 0.83 | 0.75 | 0.91 | 0.83 | 0.82 | 0.878 | Transformer |
| **GPTanam** | *0.91* | *0.96* | *0.98* | *0.97* | *0.84* | *0.97* | *0.98* | *0.94* | *0.969* | Transformer |
| **Hist Gradient Boosting** | 0.87 | 0.84 | 0.89 | 0.78 | 0.70 | 0.94 | 0.78 | 0.78 | 0.850 | Ensemble |
| **KNN** | 0.80 | 0.83 | 0.83 | 0.78 | 0.71 | 0.89 | 0.78 | 0.61 | 0.783 | Nearest Neighbor |
| **LGBM** | 0.82 | 0.90 | 0.92 | 0.79 | 0.77 | 0.91 | 0.82 | 0.81 | 0.873 | Ensemble |
| **Linear SVC** | 0.83 | 0.90 | 0.90 | 0.79 | 0.75 | 0.91 | 0.80 | 0.82 | 0.864 | Linear |
| **Logistic Regression** | 0.78 | 0.90 | 0.91 | 0.81 | 0.79 | 0.91 | 0.80 | 0.80 | 0.864 | Linear |
| **Passive Aggressive** | 0.86 | 0.89 | 0.88 | 0.78 | 0.72 | 0.91 | 0.77 | 0.79 | 0.842 | Linear |
| **Perceptron** | 0.85 | 0.89 | 0.88 | 0.78 | 0.70 | 0.91 | 0.78 | 0.80 | 0.846 | Linear |
| **Random Forest** | 0.87 | 0.91 | 0.88 | 0.80 | 0.66 | 0.93 | 0.79 | 0.66 | 0.840 | Ensemble |
| **Ridge CV** | 0.87 | 0.90 | 0.91 | 0.79 | 0.75 | 0.91 | 0.80 | 0.82 | 0.866 | Linear |
| **SGD** | 0.88 | 0.91 | 0.91 | 0.8 | 0.80 | 0.90 | 0.81 | 0.82 | 0.870 | Linear |
| **XGB** | 0.83 | 0.91 | 0.91 | 0.79 | 0.74 | 0.90 | 0.82 | 0.8 | 0.870 | Ensemble |
| **N per class (Test)** | 41 | 498 | 1554 | 186 | 112 | 568 | 985 | 371 | 4315 | |

*Table III.7 Micro F1-scores for all classes and average F1-score for all models AE: Accident of Exposure (to Bodily Fluids), MVA: Motor Vehicle Accident*

Average macro precision and micro F1-scores were systematically higher with the transformers than with the TF-IDF/SVM couple on the complete test dataset, as seen in Table III.7. Among the transformers, GPTanam achieved an average micro F1-score of 0.969, outperforming CamemBERT, FlauBERT and BelGPT2 for which F1-scores were 0.878, 0.873 and 0.887 respectively. Macro-average precision was higher than F1-score in almost all cases, except for TF-IDF/SVM where macro precision was lower than micro F1-score (macro precision = 0.860, micro F1-score = 0.864).

| Type of trauma | N | TF-IDF/ SVM | CamemBERT-CCNET | FlauBERT-small | Bel-GPT2 | GPTanam |
|---|---|---|---|---|---|---|
| **Accident of Exposure to Bodily Fluids** | 41 | 0.83 | 0.84 | 0.84 | 0.83 | **0.91** |
| **Assault** | 498 | 0.90 | 0.91 | 0.92 | 0.91 | **0.96** |
| **Fall** | 1554 | 0.90 | 0.92 | 0.91 | 0.92 | **0.98** |
| **Foreign Body in Eye** | 186 | 0.79 | 0.84 | 0.82 | 0.82 | **0.97** |
| **Intentional Injury** | 112 | 0.75 | 0.76 | 0.73 | 0.77 | **0.84** |
| **MVA** | 568 | 0.91 | 0.90 | 0.91 | 0.91 | **0.97** |
| **Other trauma** | 985 | 0.80 | 0.83 | 0.82 | 0.85 | **0.98** |
| **Sport Accident** | 371 | 0.82 | 0.83 | 0.83 | 0.85 | **0.94** |
| **Total** | 4315 | | | | | |
| **Micro F1-score** | | 0.864 | 0.878 | 0.873 | 0.887 | **0.969** |
| **Macro precision** | | 0.860 | 0.880 | 0.880 | 0.89 | **0.970** |

*Table III.8* Micro F1-scores for all classes and selected models with micro average F1-scores and macro average precision on the complete test dataset.

The distribution of *n* clinical notes per class not being balanced, the micro-F1 scores were, in all cases, lower with the classes where n was lower. Concerning the micro F1-score of the different classes, GPTanam had higher scores than the other transformers and TF-IDF. The performance of GPTanam was high for all classes except for intentional injuries. We made the assumption that these results might be associated with the semantic heterogeneity and variety of this particular class. Indeed, this class encompassed self-arm (self-mutilation, punching due to rage, self-stabbing) and suicide attempts (shooting, alcohol or drug poisoning, car crashing) with few examples per injury. On the other hand, classes such as MVA or fall have semantic consistency with larger number of examples. The confusion matrix is given in Figure III.11. An error analysis of the “intentional injury class, as well as the other classes, is provided in the next section.


*Figure III.11 Confusion matrix of GPTanam model on the full test dataset*

### III.3.6 Error analysis

Accident of exposure to bodily fluids: The unigram analysis showed that the key words "contact blood" were absent in the top 10 bigrams in the incorrectly classified clinical notes, while on the other hand unigrams analysis shows that "HIV" is the 9th unigram (after "aes", "blood", "needle", "source", "intercourse", "dakin", "work", "sexual").

Assault: Regarding the class “Assault”, the top-3 bigrams were "physical assault", "declare having", and "punch" (fr: coup poing) for the correctly classified clinical notes while "left hand", "hand trauma", "mechanical fall" were the most frequent bigrams. The verification of the 18 clinical notes manually annotated as “Assault” showed that for 11 of them the label predicted by the model was correct (1 fall, 8 self-harm, 1 MVA, 1 sport accident paintball).

MVA: The acronym “mva” (n=700) was the most represented unigram in the correctly classified corpus while “pain” was the most represented one in the clinical notes classified as not MVA. When analyzing the 6 incorrectly classified clinical notes, 3 of them were wrongly labeled as they were in fact referring to an assault, a fall and a basketball accident. The 3 remaining clinical contained two types of trauma such as falling on the street.

Foreign body in the eye: The unigram analysis for this class showed that the unigrams "eye" and "theeye" were the most represented (n=140) while "left" and "hear" were the top-2 unigrams in the clinical notes classified as not being “foreign body in the eye”. In fact, one of these clinical notes was related to a foreign body in the ear and two others were assault without mention of eye trauma.

Fall: The top-3 bigrams for the correctly classified clinical notes were "mechanical fall", "loss of consciousness", "cranial trauma" and were "right ankle", "ankle trauma", "left ankle" for the incorrectly classified ones. Twenty-one of the incorrectly classified clinical notes encompassed a double mechanism of trauma involving a sport accident, 16 MVA and 4 assault as well as a fall were present. Nine notes mentioned back pain, ankle and knee twists, pain while getting off of a truck, a patient found at the bottom of stairs, without mention of falling.

Intentional Injury: The most frequent uni and bigrams were different between the correctly and wrongly classified clinical notes. The most represented unigram and bigram were, respectively, “imv” (fr, voluntary drug intoxication) and “suicide attempt” in the correctly classified corpus of clinical notes while, “hand” and “punch given” were the most common in the correctly classified notes. Indeed, the model classified 10 clinical notes as assault while these clinical notes were related to a patient having punched something or himself.

Sport: The most frequent unigrams for correctly classified clinical notes were "pain", "left" and "trauma" and the bigrams were "right ankle", "functional impotence" and “left knee". The most frequent unigrams and bigrams for the incorrectly classified notes were, respectively “fall", "trauma", "bike" and "bike fall", "right knee", "knee pain". Thirteen falls occurred while biking without mention of the place and were classified as MVA. Five incorrectly classified notes were eye trauma while practicing sport.

Removing complex/ambiguous clinical notes is associated with an increase of performance for all models, the average gain of F1-scores being 0.04 for TF-IDF/SVM, CamemBERT, FlauBERT and BelGPT2. The average gain of micro F1-score was 0.01 for GPTanam, which seems more robust to complex and/or ambiguous content.
Difference in performance when potentially complex/ambiguous content is taken into account was greater with TF-IDF/SVM, CamemBERT, FlauBERT and BelGPT2 than with GPTanam, especially with the classes MVA and Sport Accident where the average gain of micro F1-score per class was 0.07 as seen in Figure III.10. Performance for the class "Accident of exposure to bodily fluids" did not improve for TF-IDF/SVM, CamemBERT and FlauBERT when complex/ambiguous content was removed from the test dataset. Performance of GPTanam did not improve for the GPTanam with the classes "Foreign body on the eye" and "Other trauma" but F1-scores were already very high with, respectively, F1-scores of 0.97 and 0.98. Performance was slightly improved for ’Assault’, ’Fall’, ’MVA’, ’Sport Accident’ and ’Other trauma’ when potentially complex and/or ambiguous content was removed from the test dataset for all models as seen in Table III.9 and confusion matrix in Figure III.13.

| Type of trauma | N | TF-IDF/ SVM | CamemBERT | FlauBERT | Bel-GPT2 | GPTanam |
|---|---|---|---|---|---|---|
| **Accident of Exposure to Bodily Fluids** | 36 | 0.81 | 0.84 | 0.84 | 0.84 | **0.90** |
| **Assault** | 474 | 0.93 | 0.93 | 0.94 | 0.93 | **0.97** |
| **MVA** | 541 | 0.97 | 0.97 | 0.97 | 0.98 | **0.99** |
| **Foreign Body in Eye** | 177 | 0.80 | 0.85 | 0.83 | 0.83 | **0.97** |
| **Fall** | 1348 | 0.95 | 0.96 | 0.95 | 0.97 | **0.99** |
| **Sport Accident** | 318 | 0.89 | 0.91 | 0.91 | 0.93 | **0.98** |
| **Intentional Injury** | 95 | 0.79 | 0.80 | 0.75 | 0.81 | **0.85** |
| **Other trauma** | 859 | 0.84 | 0.86 | 0.86 | 0.89 | **0.98** |
| **Total** | 3848 | | | | | |
| **Micro F1-score** | | 0.904 | 0.921 | 0.918 | 0.932 | **0.981** |
| **Macro precision** | | 0.902 | 0.921 | 0.919 | 0.932 | **0.982** |

Table III.9 Micro F1-*scores for all classes and selected models with micro average F1-scores and macro average precision on the test dataset without ambiguous content.*

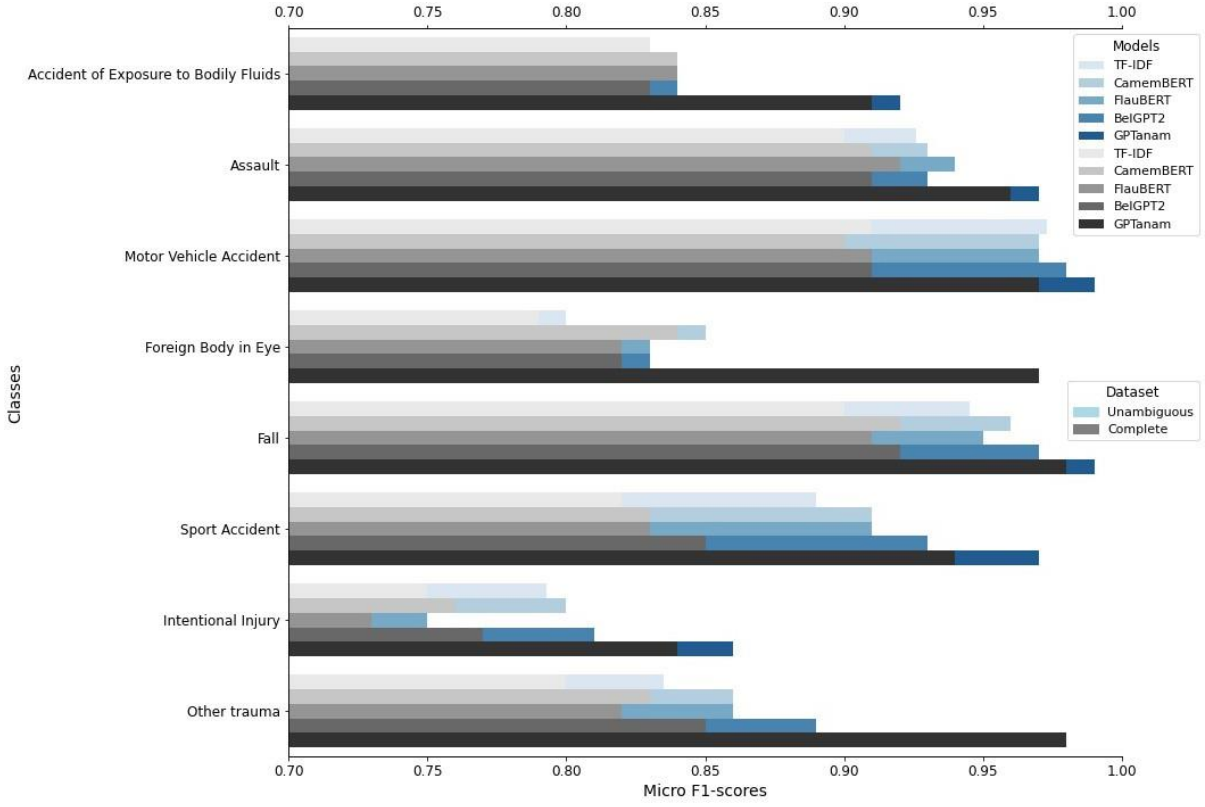



*Figure III.12 Plot of micro F1-scores of all models for each class for both the complete test dataset (blue bars) and the test dataset without potentially ambiguous content as regard to its classification (grey bars).*

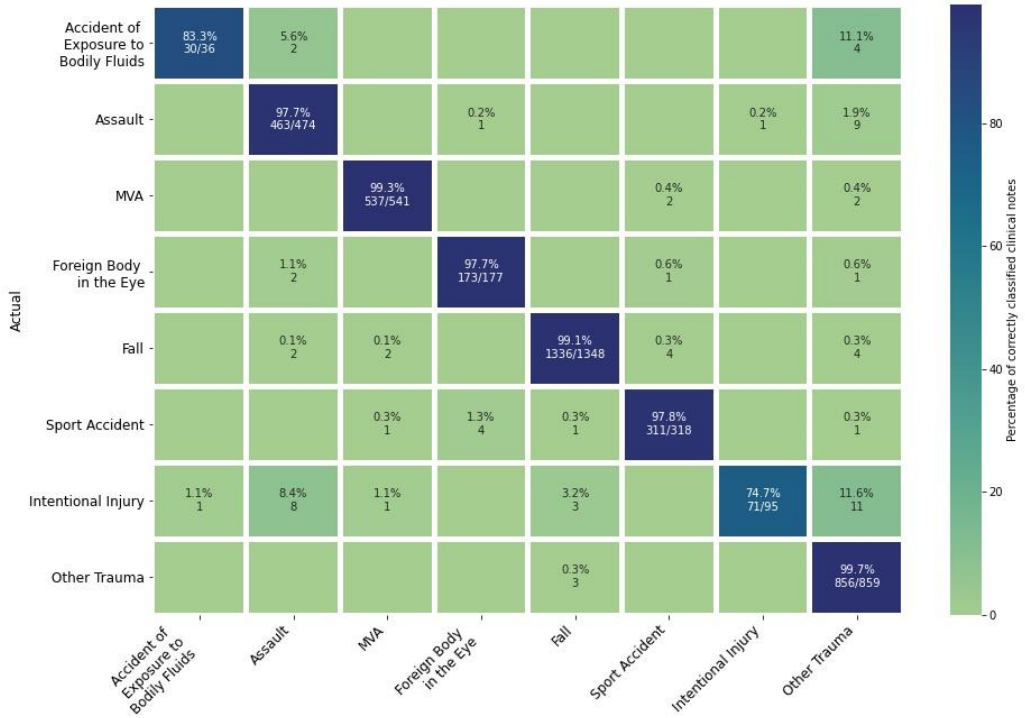



*Figure III.13 Confusion matrix of GPTanam model on the test dataset without ambiguous content*

## III.4 TARPON: Discussion

### III.4.1 Transformers: a new state of the art

The transformers applied to the free text data from the ED of the Bordeaux University Hospital showed interesting results reaching an average micro F1-score of 0.969 for a GPT2 model with a French tokenizer and with a self-supervised training step on a domain specific corpus in addition to a large French corpus. This model showed better performance than TF-IDF/SVM and the other transformer models on average metrics and on all classes. In 2018, when reviewing deep learning algorithms for clinical NLP, Wu et al. projected the rise in popularity of Transformer models.[36] However, some studies show that traditional approaches, when tailored to the specific language and structure of the text inherent to the classification task, can achieve or exceed the performance of more recent ones based on contextual embeddings such as BERT[37]. Further study could involve comparing our model's performance to Bi-LSTM with pre-trained embeddings such as Word2Vec or transformers embeddings and CNN.

### III.4.2 Self-supervised training on domain specific corpus and tokenizer

The decision to use pre-trained models on French corpora with a French tokenizer has probably contributed to the global performance of the chosen transformer models. General language transformer models pre-trained on a cross-domain text corpus in a given language have flourished recently. BelGPT2 was the first GPT2 model fine-tuned on a French heterogeneous corpus (CommonCrawl, French Wikipedia, EuroParl...) released on the hugging face platform. Self-supervised training of transformers on a specific domain can improve task performance such as classification[38], text generation[39] and predicting hospital readmission.[40] Despite lots of experiments using BERT, GPT-2 hasn't been studies as well as BERT yet. Our team showed that the number of data required to achieve a given level of performance (area under the curve over 0.95) was reduced by a factor of 10 when applying self-supervised training on emergency clinical notes to a binary classification task.[41] Here, we confirm the benefits of a self-supervised training step on a domain-specific corpus. However, it is questionable whether this approach will be applicable when extending the TARPON project to data from other EDs in France, as each region or ED uses a specific language in addition to the medical language, which uses many abbreviations that can vary locally (i.e. assault is written "brawl" in Bordeaux, "hep" means hepatitis...). A possible solution would be to train the model on a corpus resulting from the extraction of ED notes at a national level. Similarly, the treatment of medical concepts and abbreviations remains an area for improvement, as not all EDs use the same abbreviations in the same context. The use of ontologies developed in the field of emergencies could constitute an area for improvement. Transformers have also recently been tested for the identification and replacement of abbreviations with good results for BERT[42,43], however, there has not yet been a test on data from a mixture of common language and medical terms in French.

In addition, as the authors who proposed the CamemBERT model did not compare the different models from the OSCAR, CCNet and Wikipedia datasets on a classification task, a future work could compare the different sets on our database. In this logic, it would be appropriate, while we have only used the basic models of CamemBERT, FlauBERT and GPT-2, to test the different sizes of pre-training datasets on a classification task as well as the different sizes of models. Indeed, Martin's team has shown that the standard CamemBERT model (110 million parameters) trained on all 138GB of OSCAR text, does not massively outperform the model trained "only" on the 4GB sample in morphosyntactic labeling, syntactic parsing, Named Entity Recognition (NER) and Natural Language Inference (NLI)[44]. One perspective considered is to test different models of French transformers that have been released since CamemBERT, FlauBERT or BelGPT2 such as Pagnol or BARThez.

### III.4.3 Taxonomy

The performance of the models improved when we excluded the clinical notes that we considered to be the most complex and/or ambiguous from our test dataset. The classification errors analysis showed that when clinical notes encompassing two mechanisms of trauma (i.e. "fall from bike on the street") were removed from the test dataset, models performed better. This expected result shows that since the advent of transformers, the margin of progress in a free text classification task is nowadays low. This behavior was less important with GPTanam, which

seems to have benefited from the self-supervised pre-training phase for reducing classification errors by learning semantic representations beforehand. However, the annotation grid created for the project is partly responsible for some classification errors in the sense that there are areas of semantic overlap between classes. In addition, the coding system used did not allow for the coding of several traumatic mechanisms (e.g., a collision between two individuals, followed by a fall). To be able to account for these situations, a new coding system will be used for the next phases of the project, using the recently released version of trauma classification grid used by the FEDORU (Fédération des Observatoires Régionaux des Urgences) and OSCOUR.

## III.5 TARPON: Conclusion

Transformers have shown great effectiveness in a multi-class classification task on complex data encompassing narrative, medical data and jargon. The choice of this type of architecture in the automatic processing of emergency department summaries in order to create a national observatory is relevant. Applying a self-supervised training step on a specific domain corpus has substantially improved classification performances with a French GPT2 model.

## III.6 TARPON: Perspectives

The first phase of the TARPON project has enabled:

- The creation of a first annotation grid (Appendix J and Appendix K)
- The development and validation of a Transformer model for the classification of clinical notes written by ED professionals from a single hospital's EHR[243].
- The application of this model for calls to the EMD in Gironde department[244].
- The creation of a de-identification algorithm using a Transformer[31].

### III.6.1 Improvement of the annotation grid

Firstly, the annotation grid developed for the TARPON project did not allow for the manual labeling of several traumatic mechanisms (e.g., a collision between two individuals, followed by a fall) resulting in ambiguous classifications for both the annotators and our model. Furthermore, multiclass classification had not been envisaged in the labeling strategy which led to a unique class attributed to a given medical record. The underlying reason was that in the daily reporting hosted by the Oscour Network, one and only trauma mechanism would be added. To overcome this shortcoming, an improved annotation grid was discussed.

Several strategies were discussed, and the adoption of the International Classification of External Causes of Injury[35] seemed be the best option for benefiting a standardized classification tool in line with WHO recommandations[5]. Hence, to improve ergonomics and accuracy, we translated the entire ICECI into French and created training material and a dedicated annotation software.
The ICECI was related to the External Causes chapter of the ICD-10[245]. Both the ICECI and the External Causes chapter of the ICD-10 provided ways to classify and code external causes of injuries. Different design criteria have resulted in considerable differences between the two

systems, and comprehensive mapping at fine level was not possible. In 2022, the WHO announced that the ICECI was no longer maintained. Indeed, the experience with ICECI had informed the redesign of the relevant chapter of the ICD-11 and the different elements of ICECI had been included as extension codes in the ICD-11. The ICD-11 was released in 2018 and endorsed by WHO in 2019[246]. To date, the disease classification used in France is still ICD-10; ICD-11 has not yet been translated or validated.

Meanwhile, in 2021, the FEDORU (Fédération des Observatoires Régionaux des Urgences) set up the ground rules for the improvement of the emergency field data collection. The objectives of this work, carried out by FEDORU, are multiple and include the extension of data collection to EMD (SAMU, SMUR), technical platforms and downstream services, but also the increase in the frequency of data transmission (every 5 minutes) and the modification of the RPU format[247]. Some key elements of the RPU such as the circumstances and the setting of occurrence as well as the patients' acuity (with the triage score) are planned to be added in the future version of the RPUs (v3). Based on the experience with the TARPON annotation grid and the adaptation of the ICECI, we were able to collaborate with the FEDORU to propose a thesaurus for the circumstances of the event (which include the mechanisms of trauma) and the chiefs of complaint. We also proposed the addition of items such as activity during the event, intentionality, antagonist in the case of a traffic accident and suspected alcohol and/or drug use in order to follow the WHO guidelines for injury surveillance systems.

### III.6.2 Future epidemiological steps of the TARPON project

The validation of the trauma classification tool made it possible to assign labels to the entire emergency database of the Bordeaux University Hospital. These labels, derived from the initial annotation grid, allow a precise description of each trauma by associating concepts such as pre-traumatic faint, suspicion of alcohol or drug consumption and, in the case of public road accidents, the means of transport, the counterpart and protective equipment.
Thanks to this comprehensive database, studies are being or will be carried out on road traffic accidents, triage and, more importantly, medical factors (such as drug use and medical conditions) of exposure to trauma risks. This last project is being carried out in collaboration with the Health Data Hub, which allows our database to be linked to the SNDS (Système National des Données de Santé), and its main objective is to demonstrate the possibility of carrying out epidemiological studies using automatic labelling.

### III.6.3 Towards a French trauma surveillance system

As the proof of concept for the clinical note's classification tool has been validated, further steps are needed before moving to a national scale. As mentioned in section 19II, clinical notes' structure (length, framework) and language can vary across professional categories, genders, social groups, personalities, hospitals, areas and regions. Hence, the heterogeneity of all clinical notes across France must be accounted for. The second phase of the TARPON project aims to collect ED databases from diverse regions and types of hospitals to evaluate the pipeline or enhance it. So far, about 10 hospitals agreed to participate (Agen, Arcachon, Blayes, Langon, Libourne, Limoges, Mont de Marsan, Pau, Poitiers and Bordeaux Saint-André ED). In collaboration with the FEDORU, a standard agreement has been drawn up to regulate

the hosting of files on the project server. It has already been signed by the Regional Hospital of Agen and will then have to be sent to all the other partner sites.

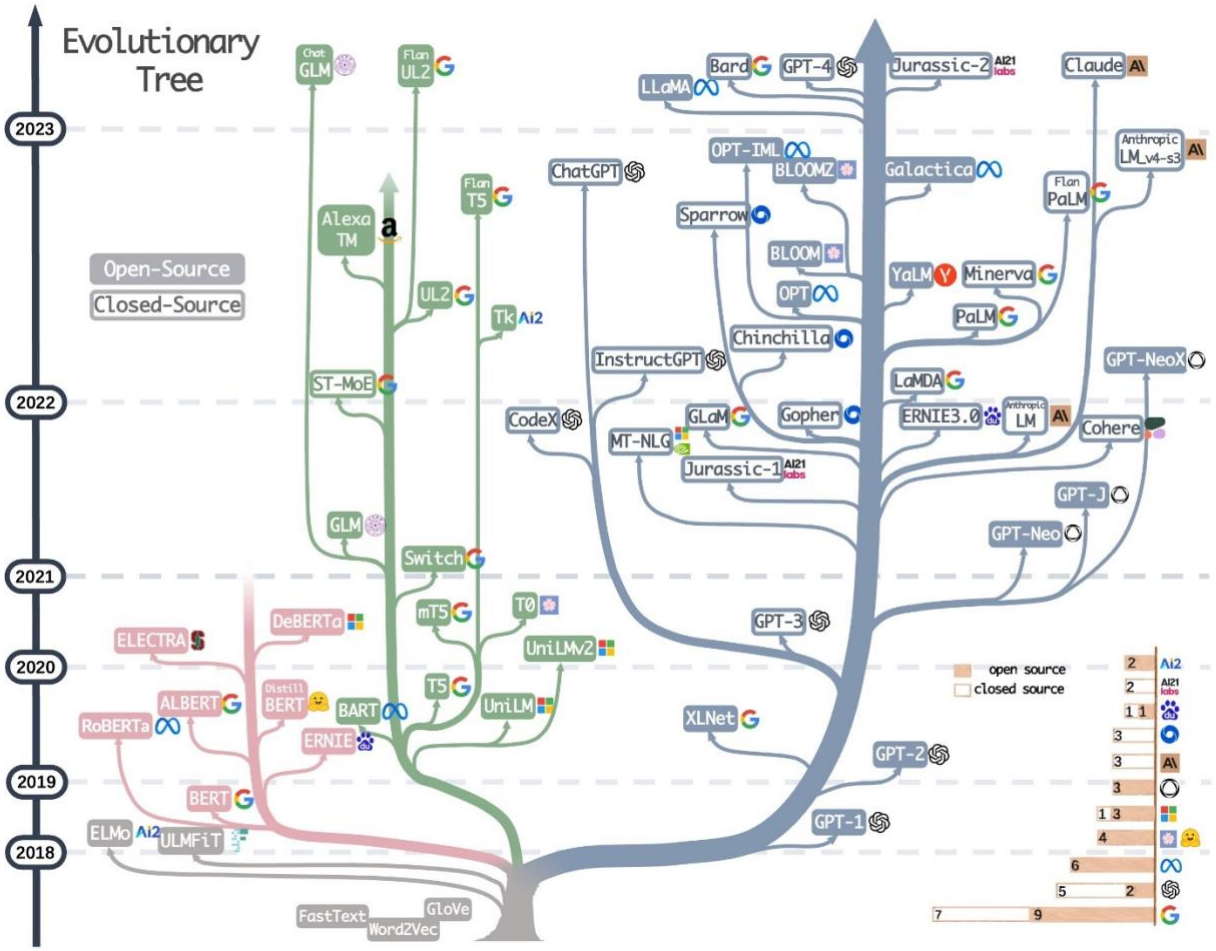



*Figure III.14 From "Harnessing the Power of LLMs in Practice: A Survey of ChatGPT and Beyond"[218] The evolutionary tree of modern LLMs traces the development of language models in recent years and highlights some of the most well-known models.*

*Models on the same branch have closer relationships. Transformer-based models are shown in non-grey colors: decoder-only models in the blue branch, encoder-only models in the pink branch, and encoder-decoder models in the green branch. The vertical position of the models on the timeline represents their release dates. Open-source models are represented by solid squares, while closed-source models are represented by hollow ones. The stacked bar plot in the bottom right corner shows the number of models from various companies and institutions.*

The challenge of the increase in heterogeneity as a result of the aggregation of different databases could be addressed by the Large Language Models (LLMs) currently being proposed. As the development and open-sourcing of LLMs has increased exponentially in recent years and months, as can be seen in Figure III.14, the possibility of performance improvement for our model increases all the time. However, the use of this type of model raises technical, ethical and alignment issues. We propose to go further by raising the level of reflection to the use of AI in emergency medicine in the following section.

# IV. ARTIFICIAL INTELLIGENCE IN EMERGENCY MEDICINE: VIEWPOINT OF CURRENT APPLICATIONS AND FORESEEABLE OPPORTUNITIES AND CHALLENGES

## IV.1 Flow Challenges in Emergency Departments

Emergency Departments (EDs) and related services such as Intensive Care Units and Emergency Medical Dispatch (EMD) have recently been in the spotlight due to the covid-19 pandemic. Overcrowded services, very long waiting times, and staff exhaustion have highlighted a fragile system with difficulties in responding to such exceptional situations. Even in period with routine activity levels, waiting times, and the optimization of the patient healthcare pathway have already been the subject of national efforts in France, pointing the need to rethink the emergency system. Indeed, the number of ED visits worldwide has increased faster than the rate of population growth in the past decades[248–250].

### IV.1.1 Factors of Emergency Departments crowding

The identified causes of increasing ED attendance include non-urgent visits, frequent visitors, extending boarding times, staff shortages, and repeated reductions of downstream beds[251]. Using the conceptual model of ED crowding developed by Asplin et al.[252], which divides ED crowding into three interdependent components, the causes of crowding can be broadly categorized as identifying input, throughput or output causes:

- Input: Causes of crowding related to the input phase of the ED process suggest increases in patients' venues with urgent and complex needs[253–256] , with low-acuity chiefs of complaints[257], or represented by the elderly [253,255,258,259] , as the main factors. Access to appropriate care outside of the ED has also been identified as an issue as well [255,257,260]

- Throughput: Internal factors of ED crowding identified are ED nursing staff shortages [261,262], delays in receiving laboratory test results and delays in patient management decisions [263.]

- Output: All studies that reported on output factors as a cause of ED crowding concluded that access block, that is, the inability to transfer a patient out of the ED to an inpatient bed once their ED treatment has been completed, was the major contributor[263–268].

### IV.1.2 Consequences of Emergency Departments crowding

The negative effects of ED crowding include impact on several patient oriented outcomes such as mortality [249,269,270], complication rates[248] , walkouts[271] , time to treatment [248,272], satisfaction[273], and length of stay[274]. Furthermore, ED crowding has been pointed as a major stress factor for healthcare professionals leading to burnouts[275] and medical errors[276]. So far, solutions and efforts have mainly focused on improving patient workflow within the ED, however, a more comprehensive approach appear more effective[277].

## IV.2 ARTIFICIAL INTELLIGENCE : A POSSIBLE SOLUTION

### IV.2.1 Artificial Intelligence in Emergency Medicine: Current Applications and Foreseeable Opportunities

The field of emergency medicine has received considerable interest in the application of AI to health care owing to the unique nature of this medical practice. With challenges related to organization and coordination as well as the need for rapid and accurate decision-making for patients categorized as high acuity, novel approaches provided by AI are promising in emergency medicine and services. AI techniques have already been shown to be promising for improving diagnosis, imaging interpretation, triage, and medical decision-making within an ED setting[278]. However, most research on AI in emergency medicine is retrospective and has not led to applications beyond the proof of concept. Therefore, the potential for AI applications in routine clinical care settings is yet to be achieved. Critical appraisal of evidence supporting whether a clinical digital solution involving AI has an impact on patient outcomes should be mandatory[279]. Specifically, an independent evaluation by an objective independent entity (or authorized entities), both during development and use, should be performed. The independent evaluation would address verification, validation, and impact on patient outcomes and safety. To date, few system suppliers have challenged their products and services in terms of key health metrics[280]. However, some applications have already been deployed for prehospital, EMD, and ED (Figure IV.*1*). In this contribution, we attempt to depict the landscape of AI-based applications currently used in the daily emergency field. For each section, we will provide a context based on recent reviews, the AI applications' algorithms or models used (if available), how they were validated, and whether the desired impact on patients' outcomes was assessed. We also propose future directions and perspectives. Our second objective is to examine the legal and ethical specificities of AI use in the emergency field.

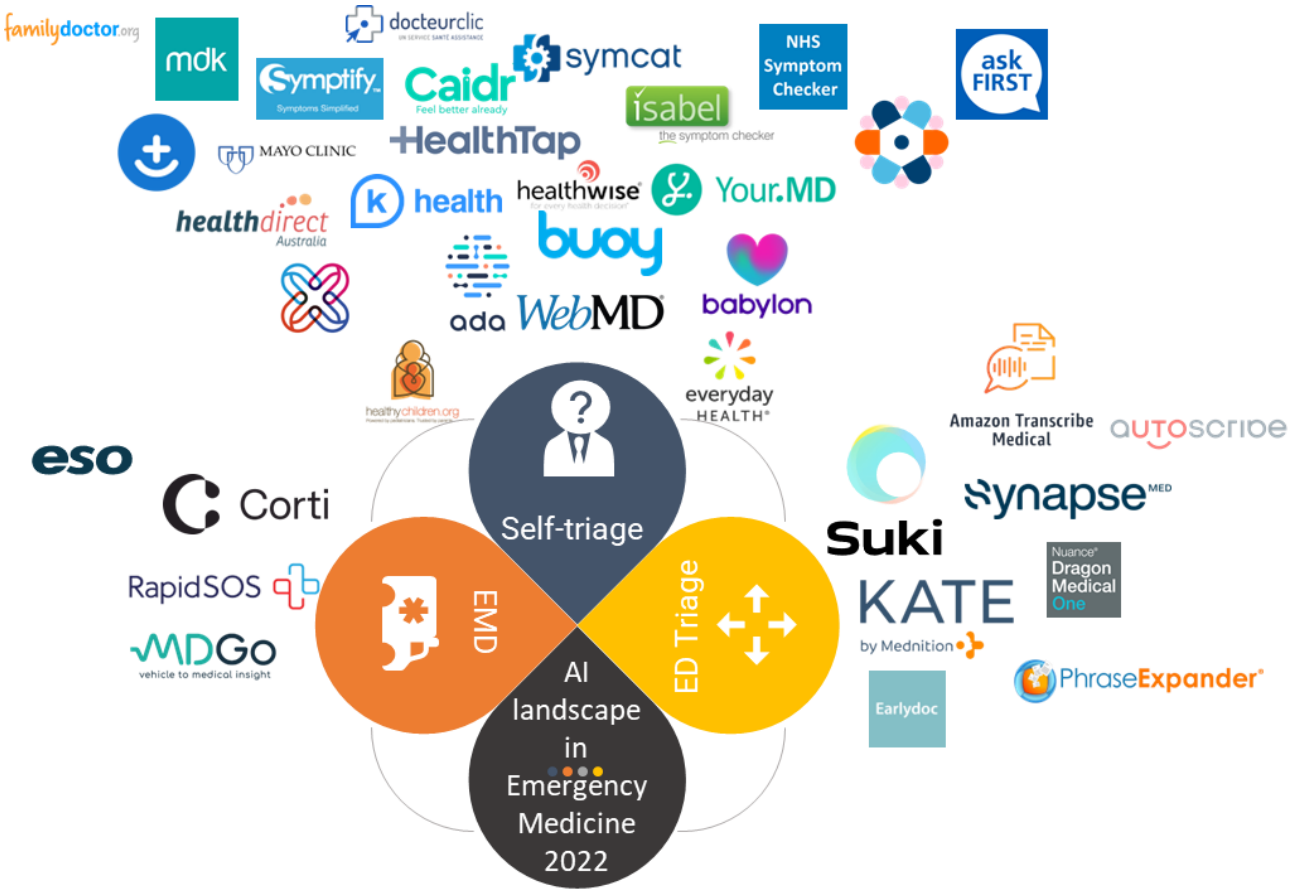



Figure IV.1 AI business Landscape in Emergency Medicine in 2022

The journey of a patient who requires care in the ED includes several steps that can or could be impacted by AI Figure IV.*2*. Before coming to an ED, several steps can be carried out such as checking symptoms on the internet and contacting the emergency call center or their general practitioner.

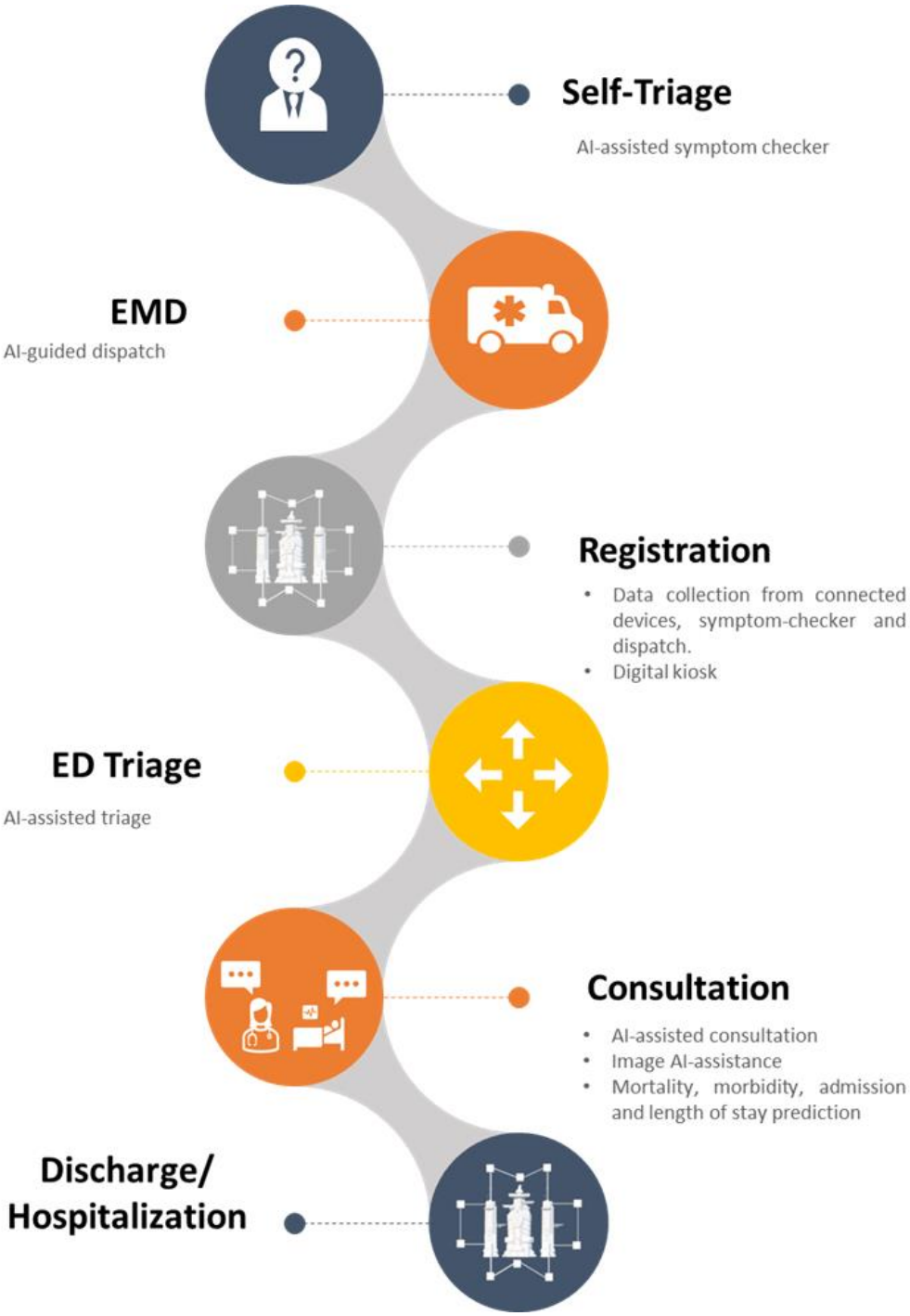



Figure IV.2 The emergency patient journey and where Artificial Intelligence is making or can make an impact.

AI: artificial intelligence; ED: emergency department; EMD: emergency medical dispatch.

#### *IV.2.1.1. Pre-hospital*

Self-Triage

The use of patient-facing clinical decision support systems (CDSSs) has continuously increased in recent years years[281]. Tools assisting laypersons in their self-assessment of whether and where to seek urgent professional medical care and for what diagnoses based on the users' input of symptoms and medical history are termed symptom checkers. To date, symptom checkers provided by free websites or mobile apps have proven to be inconsistent, supplying generally risk-averse advice and often recommending more urgent care than necessary necessary[282,283]. Digital tools that impact care delivery and behaviors should undergo rigorous evaluation that enables evidence-based determination of their efficacy. However, evaluations of the effectiveness of self-sorting apps often provide limited evidence as they rely heavily on observational studies[284]. Schmieding et al[285] recently assessed the triage accuracy of 22 symptom checkers and showed that their performance did not improve between 2015 and 2020. For 2 cases of use, the triage performance decreased (advice on when emergency care is needed and when no health care is required for the moment). The apps sample of 2020 less frequently mistook self-care cases and nonemergency cases for emergencies; at the same time, it more often misclassified emergencies as non-emergencies[285]. Regarding the algorithms or models used by these proprietary websites or apps, information about their architecture, development, and validation is sparse. When the information is available, most symptom checkers and their decision support systems rely on probabilistic or graphical algorithms (Bayesian decision trees or Bayesian-directed graphs[286–291]). Some apps, such as Babylon Health[292], use a chatbot that presents the user with unique or multiple-choice questions for symptom assessment[293]. Although there is no clear explanation of the algorithm used by Babylon, the team has released open-sourced Neural Temporal Point Processed models[294], which are integrated into an encoder-decoder framework based on deep learning. This indicates that the app likely uses this type of model[295]. To ensure the safety of symptom checker users, transparency about the algorithms used should be maintained. Further research and development also seem necessary for improving these self-sorting tools. The use of deep learning models for these apps should be considered to attempt improving their limited efficacy (Textbox IV.*1*).

- Multiple proprietary self-sorting apps
- Lack of validation studies
- Weak evidence for their efficacy
- Algorithms often undisclosed

Textbox IV.1 Summary of pre-hospital clinical decision support systems' assessment

EMD

Pre-hospital emergency care and ambulance demands have significantly increased over the past decade[296–298]. Emergency medical dispatch involves the receipt and management of demands for urgent medical assistance. It encompasses 2 main dimensions:

call answering, where emergency medical calls are received and events are classified according to their priority (triaged) and coordinating, where the best available resources are dispatched to manage the event.

### EMD Data Entry

Emergency medical dispatchers at EMD centers play a pivotal role in coordinating prehospital care. The interaction between the dispatcher and patient results in documentation that can be guided (structured form), semiguided (semistructured), or free (unstructured). Although effective in narrow and predictable domains, structured data entry can be quite slow when events are wide ranging and heterogeneous. To address this issue, the already-in-use Corti[299] system assists emergency dispatchers by analyzing the caller's speech and description. This system provides advice on which questions to ask next, indicating when a patient may have a particular presentation, such as myocardial infarction or stroke. It also helps in data extraction, where the system can extract and pull information on the caller's address and location to reduce the time needed to complete the call and dispatch emergency medical services. The framework of Corti contains 2 models: an automatic speech recognition (ASR) model that transcribes speech to text and an out of hospital cardiac arrest (OHCA) detection model that predicts OHCA events from transcribed speech in real time. The ASR is a deep neural network using a model based on Connectionist Temporal Classification[300]. This end-to-end (E2E) deep learning framework is based on a recurrent neural network, and the network outputs are transformed into a conditional probability distribution over label sequences (letters, words, or sentences of the caller). The network can then be used as a classifier by selecting the most probable label for a given input sequence [38]. For each second of raw audio, the classifier predicts whether there is an OHCA based on the accumulated audio sequence[299]. The efficacy of the AI-guided system provided by Corti was assessed for OHCA by Byrsell et al, and it was shown that the E2E model recognized OHCA faster than dispatchers[299]. Despite the promising results for OHCA, the study assessing the system was retrospective, and other critical conditions were not tested.

Semistructured or free-structured text observations are the most frequently used input format for EMD, according to Miller et al[301]. If dispatchers require this format to be continued in the future, solutions to facilitate, speed up, and optimize this type of input should be considered. Computed free text involves natural language processing (NLP), and a recent breakthrough revolutionized this area in 2018 when the Transformer architecture was introduced by Vaswani et al[66] in "Attention is all you need." The Transformer aims to solve sequence-to-sequence tasks while easily handling long-range dependencies (problems for which the desired output depends on inputs presented at times far in the past). It relies entirely on self-attention to compute its input and output representations without using sequence-aligned recurrent neural networks or convolutions. The Transformer architecture has evolved, and some models such as the Bidirectional Encoder Representations from Transformers[189] and the Generative Pretrained Transformer 2[189] have achieved unprecedented performances on various NLP tasks such as classification, question answering, named entity-recognition, relation-extraction, or sentence-similarity tasks [183,302]. A major efficient feature of Transformers that dispatchers could benefit from is text generation through autocompletion[42,60]. By proposing a text complement fitting the string of characters that the dispatcher would have started to type, the autocomplete would allow to speed up the typing process and thus save time for the dispatcher. The autocomplete would also limit

typing errors by entering the characters that remain to be typed without human intervention. Finally, the autocomplete would avoid the dispatcher having to correct their typing errors if necessary.

### EMD Call Waiting Time

EMD calls can increase drastically under exceptional circumstances such as mass shooting, wildfires, or when it is recommended to call the center before seeking care (eg, COVID-19) [303,304]. To reduce the waiting time before reaching a dispatcher for very acute patients in ordinary and exceptional situations, some solutions such as prioritized queue with the help of an ASR model and a classifier are starting to be considered and designed[305]. To the best of our knowledge, such solutions have not been tested or even developed yet.

### EMD Triage and ambulance dispatch

A large proportion of prehospital deaths when emergency medical services are involved are preventable, with 4.9% to 11.3% potentially preventable deaths and 25.8% to 42.7% definitely preventable deaths, as shown by Pfeifer et al[306]. The most frequent reasons evoked in this systematic review were delayed treatment of patients with trauma (27%-58%), management errors (40%-60%), and treatment errors (50%-76.6%)[306]. Treatment delays and caller management are often the result of dispatch algorithms that provide triage of patients categorized as high acuity for critical care and patients categorized as low acuity for diversion or nonurgent transport. Most of the current dispatch algorithms are rule based or encompass a human review of rule-based algorithms[301]. To date, 2 retrospective studies have shown that statistical machine learning and deep learning can improve or outperform rule-based algorithms[307,308]. Further validation and impact studies are needed to improve the current dysfunctional EMD triage, and AI should be considered for enhancing the dispatch algorithms. Start-up companies are making proposals to help reduce response times and ensure data transmission from connected devices before or during calls. For example, the RapidSOS system is an emergency response data platform that securely links data from connected devices and sensors directly to first responders during emergencies. Another promising system provided by the Israeli start-up MDGo is the use of advanced AI technology to help dispatchers know if a car accident requires an ambulance. When a car crash occurs, the system creates a medical report in real time with data regarding the forces applied on the passenger (eg, duration, moment, and vector). These data are sent automatically to the Israeli emergency medical services.

#### *IV.2.1.2. Emergency Departments*

- **Registration**

Whether generated from a symptom checker with a self-triage step, from a call to an EMD center, or a connected device, all collected data concerning patients could benefit EDs. Linking emergency medical services to ED data allows a continuum of care assessment and improvement in patient outcomes[309]. Concerns regarding interoperability,

security, accurate patient match algorithms, and the reliability of wireless networks as potential barriers to adoption were identified in a review conducted by Martin et al[310]. Several studies have demonstrated the feasibility of various statistical models for electronic health record (EHR) linking with EMD systems[310]. For example, Redfield et al[311] used logistic regression to link Boston's EMD electronic patient care reports with their hospital EHR and achieved an unprecedented success rate of linkage without manual review (99.4% sensitivity). The next few years will likely reveal an expansion in the use of these techniques in new ways. For patients arriving at the ED by their own means, an initial medical screening could be performed by asking a small number of questions using a smartphone or a digital kiosk set up at the ED entrance. To date, all trials entailing the redirection of patients categorized as low acuity within EDs involved human intervention and were unsuccessful or discontinued owing to adverse public relations incidents[277,312]. In a fully digitalized world, the acceptance of such solutions accompanied by awareness campaigns should be more substantial.

**ED Triage**

The check-in desk at the entrance of the ED is the first point of contact for a patient requiring emergency care where administrative agents open a specific section of the EHR. The patient then becomes a future occupant of the ED room or cubicle after being assessed by the triage nurse. Triage is a sorting process in which the "triage nurse" is required to quickly assess a large number of patients to decide the urgency of their condition and the location in the ED in which they will be evaluated and treated. Triage includes the attribution of a triage score to each patient, and several scales have been developed worldwide, with no evidence of superiority for one of them[313,314]. Even with the adoption of 5-level triage scales, the assessment still relies heavily on the subjective judgment of the triage nurse, which is subject to significant variation[315]. Furthermore, Hinson et al[316], in their systematic review, found several studies reporting low sensitivity (<80%) in identifying patients who had critical illness outcomes or died during the hospitalization. To address the lack of accuracy in the triage process, several AI-based solutions have been tested, and the authors found that there was an improvement in the health care professionals' decision-making, thereby leading to better clinical management and patient outcomes[317,318]. However, these solutions were not dedicated to triage but outcomes such as hospital admissions, mortality, or ED length of stay. An example of a real-time AI application that is already used in 16 US hospitals is provided by KATE[236,319]. Unlike most proprietary software, a validation study has been published that showed that KATE's accuracy using an extreme gradient boosting model[236] (explained on Figure IV.*3*) was 27% (P<.001) higher than the average nurse accuracy. However, no impact study has yet been published.

**Gradient boosting**

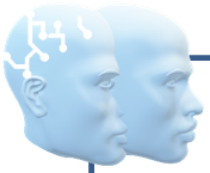
Is a special case of boosting where the errors are minimized by the gradient descent algorithm. Boosting refers to an ensemble method, which consists of training multiple models using the same learning algorithm while correcting the n-1 model's errors. Gradient compensates for the errors committed previously without deteriorating the predictions that were correct.

Figure IV.3 Gradient boosting explanation

Similar to dispatchers, the documentation workload of triage nurses can benefit from AI applications. Health care professionals currently spend up to 50% of their time documenting information in EHR[320–322]. The time spent performing documentation tasks induces both poor and inconsistent data, which may impact the quality of care[323,324]. Physicians prefer using free text over restrictive structured forms, but clinical notes often lack readability owing to an overload of acronyms and jargon[37,43], which leads to noisy, ambiguous, and incomplete data. A first improvement lever could be autocompletion, which combines automatic annotation with labels of clinical concepts. Greenbaum et al[325] and Gopinath et al[42] set up the foundations of such technologies. The Massachusetts Institute of Technology clinical machine learning group, led by Gopinath et al[42], developed a tool called Medknowts that aims to autocomplete clinical terms in the EHR while note-taking. This tool was assessed in a real ED environment and showed a 67% reduction in the keystroke burden of clinical concepts[326]. The model used is fully disclosed and is based on a shallow dual branch neural network for a minimal latency (time taken to process 1 unit of data) of approximately 0.2 milliseconds. In addition, MedKnowts allows the retrieval and display of context-specific information from a patient's EHR while unifying the documentation and search process[326]. However, the language aimed to be autocompleted with these systems is strictly medical and does not reflect the reality of clinical notes containing both nonmedical and medical concepts. Using new NLP deep learning models such as Transformers, as mentioned previously, can help handle the complexity of these type of data. Transformers have reached a state-of-the-art status for ASR by reducing the word error rate to <5 (the lower the better) on several libraries and languages[327]. Nonetheless, some challenges remain to be addressed such as latency, streaming, and adaptation capabilities for implementing E2E models. The growing progression in the technological capabilities of hospitals (servers and graphics cards) will allow for real-time efficiency without affecting the workflow. Another solution is to retrieve relevant information from real-time dialogues between health care professionals and patients. Ideally, the system would write down information in free-text form but would also extract entities such as symptoms or medications and predict scores, risk factors, and diagnosis. Vocal AI assistants such as Suki[328] and Dragon Medical One[329] are already available for health care practitioners, claiming a documentation time reduction of 72%. So far, no peer-reviewed derivation or validation studies have been found to support the legitimacy of these solutions' commercial claims.

### The Digital Hospital Concept

A digital hospital concept in the image of the digital twin[330] (Figure IV.*4*) would allow real-time bed availability. The admission and discharge data, currently collected by the admissions departments, could be transferred to the digital hospital, and the estimation of the projected bed availability rate could be made available in each department. Traditional models estimating length of stay are mostly statistical[331] or based on machine learning using the previous length of stay as input. The digital hospital model would be based on the same foundation and would also be adjusted regularly owing to a trend toward shorter lengths of stay and a shift to ambulatory medicine. The model would also be able to adjust to external data such as environmental and epidemiological factors (e.g., epidemics) in real time. Thus, if visibility on downstream beds is guaranteed, not only can waiting time in the ED be reduced when hospitalization is needed, but transfers to downstream services can also be facilitated in the event of congestion. Creating a network of all digital hospitals at the regional or state level could ensure the availability and visibility of beds and facilitate transfers between health care facilities. On a comprehensive scale, these data can provide real-time visibility of foreseeable ED arrivals and allow resources to be adapted accordingly.



**The human digital twin**

is a near–real-time copy or counterpart in cyber space of a real person in our physical world. It is one's digital description in the digital manner in a computer or a server in the cloud. When the information about the real human changes, the records change accordingly.

Figure IV.4 The human digital twin

### Improving the Patient's Waiting Time Experience

Patient experience or satisfaction with ED care is a growing area of research, and the literature has demonstrated a correlation between high overall patient experience and improved patient outcomes, cost-effectiveness, and other health care system goals[332–334]. Several factors lead to better patient satisfaction in emergency medicine such as actual waiting times[335], perceived waiting times, staff-patient communication, and staff empathy and compassion[336].

Waiting time to care in ED is the cumulative result of the time from registration assessment and the time from assessment to the initiation of medical care. This waiting time is modulated by triage in EDs when dedicated triage staff are available. Inadequate staffing has been identified as a major throughput factor associated with longer waiting times[251]. Apart from alleviating documentation tasks and facilitating flow management in ED, AI cannot propose solutions when political decisions or executives regulate staff quotas. In contrast, perceived waiting time could benefit from innovation. Waiting without information provided about delays can be a tedious and frustrating experience among people seeking urgent care, and

lack of information magnifies patients' sense of uncertainty and increases their psychological distress sometimes, leading to violent behaviors[337,338]. Transparency is a major determinant of patient satisfaction related to waiting time[275,308]. Patients provided with written or gamified ED processes tend to have a higher level of satisfaction[339,340.] Information about the estimated waiting time is provided by triage nurses or signboards at the admission desk in some hospitals. However, it has been shown that this information is not given for most patients[341]. Accurate waiting time for patients can be derived from the digital hospital with a dashboard of available places and beds. A screen indicating the waiting time in real time can be installed in the waiting room[342]. Additional information such as major events impacting the waiting time could be displayed on the screen (e.g., a pileup on the highway), and mobilizing the patient's empathy could reduce self-centered perception of care[343]. Patient-specific information on personalized waiting time estimates can also be provided via a mobile app. A positive environment can also improve a patient's perception of waiting time[344]. Distracting activities such as the use of personal cell phones can be difficult for some patients in ED rooms. The benefits of virtual reality glasses have already been demonstrated in pain management[345] and in the reduction of preoperative anxiety[346]. Hence, virtual reality glasses can also be proposed for distraction and counseling.

### IV.2.2 ED and EMD data processing enhanced by AI for public health surveillance

EDs and EMD centers generate a large volume of diverse health-related data. For public health surveillance aims, these data are most often used retrospectively and by sampling hospitals[346]. Some near–real-time surveillance systems use information extracted from EHR in addition to manual implementation provided by health care professionals[347]. These non-exhaustive procedures are time and resource consuming and are mostly based on voluntary work. Automatic signal extraction from EHR would allow real-time monitoring and ensure the responsiveness sought in any surveillance system[244,348]. The use of new state-of-the-art NLP models such as Transformers would bypass the difficulties in extracting fine-grained and standardized data from the most frequently used entries (free text) in ED and EMD.

Furthermore, with the appropriate network infrastructure, data should be collected and analyzed in real time, enabling early, accurate, and reliable signals of health anomalies and disease outbreaks. In addition, AI provides an opportunity to use various new or underexploited data sources for public health surveillance purposes, particularly those not originally or intentionally designed to answer epidemiological questions. A large amount of nontraditional data is self-generated by the public through their ubiquitous use of smart devices and social media. Public health has the potential to use real-time longitudinal data collected for health surveillance[349].

## IV.3 CHALLENGES POSED BY ARTIFICIAL INTELLIGENCE FOR EMERGENCY MEDICINE AND PUBLIC HEALTH SURVEILLANCE

### IV.3.1 Ethical and legal challenges posed by the implementation of Artificial Intelligence in Emergency Medicine

Despite the potential of AI for improving emergency clinical care, numerous ethical and legal challenges prevail. An ethical principle is a statement of a duty or a responsibility and when applied to AI technologies for health, it covers their lifecycle (Figure IV.*5* and Figure IV.*6*).

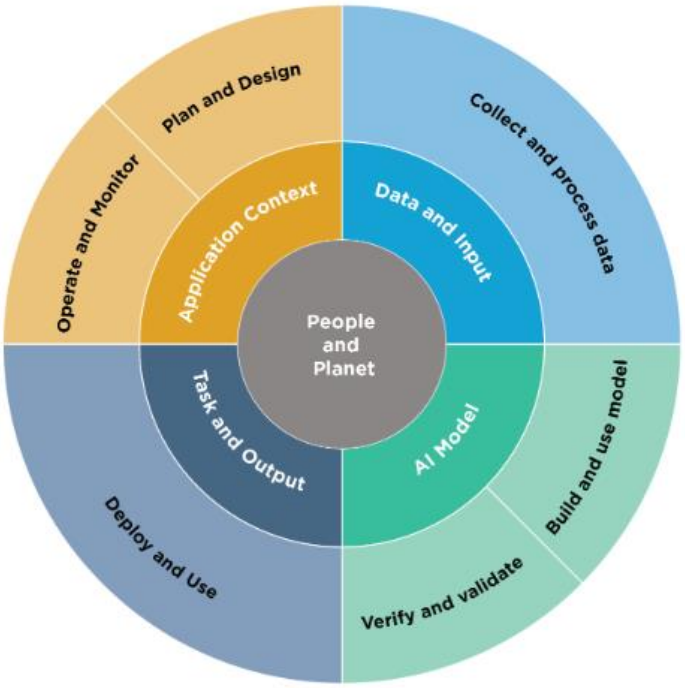



Figure IV.5 Lifecycle and Key Dimensions of an AI System. National Institute of Standards and Technology (NIST) [299]

A trustworthy AI is safe, fair and biased, is managed, transparent and accountable, explainable and interpretable, it protects human autonomy, and is privacy-enhanced[350,351]. A sense of common responsibility among all the actors involved in an AI lifecycle should prevail and healthcare providers have a special duty to adhere to these requirements because of patients' dependence on their care, should AI systems be used to assist healthcare practitioners in clinical decision-making[352]. In order to lay the foundations of trustworthy AI in emergency medicine, the ethical considerations cannot be dissociated from the legal answers that are or will be provided.

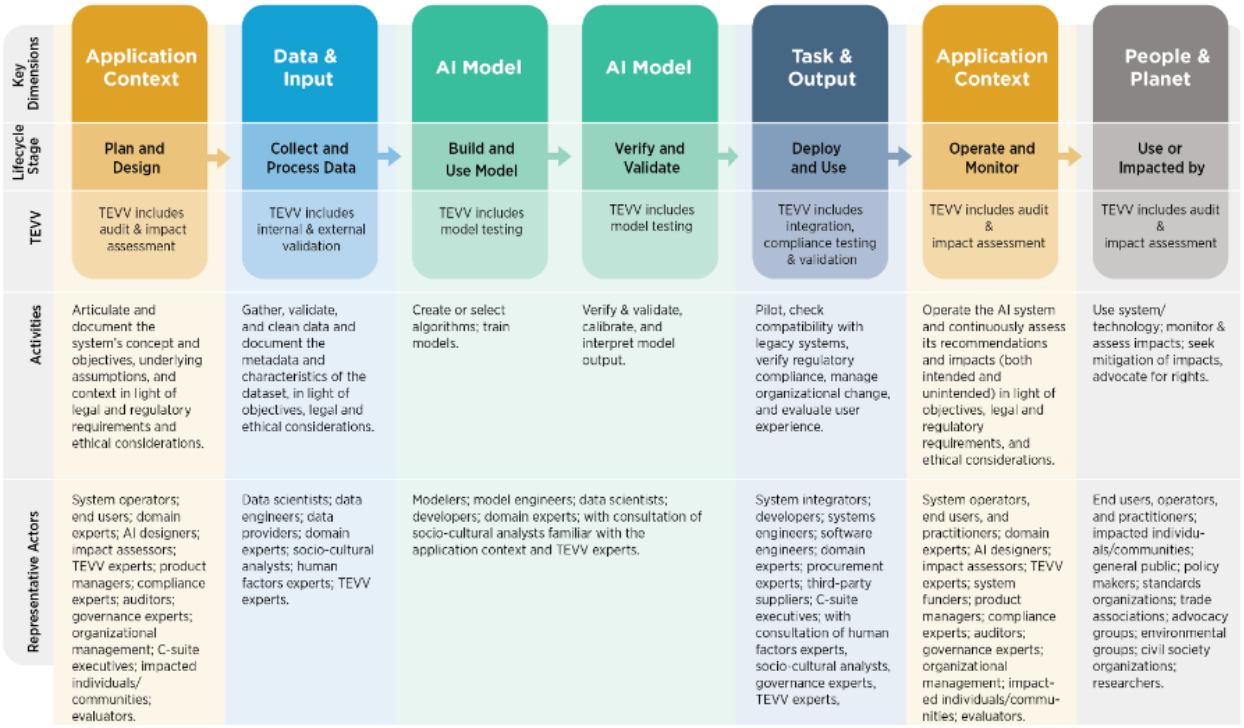



Figure IV.6 AI actors across AI lifecycle stages. Note that AI actors in the AI Model dimension are separated as a best practice, with those building and using the models separated from those verifying and validating the models. TEVV: Test, Evaluation, Verification and Validation

### IV.3.2 Safety, fairness and bias management

AI systems "should not, under defined conditions, cause physical or psychological harm or lead to a state in which human life, health, property, or the environment is endangered"[353]. Identifying, mitigating, and minimizing risks and potential harms associated with AI applications, especially in emergency medicine, are essential steps towards the development of safe AI systems and their appropriate and responsible use. Addressing AI risks and bias prospectively and continuously throughout the AI lifecycle aims at preventing misalignment (Figure IV.7)[354,355].

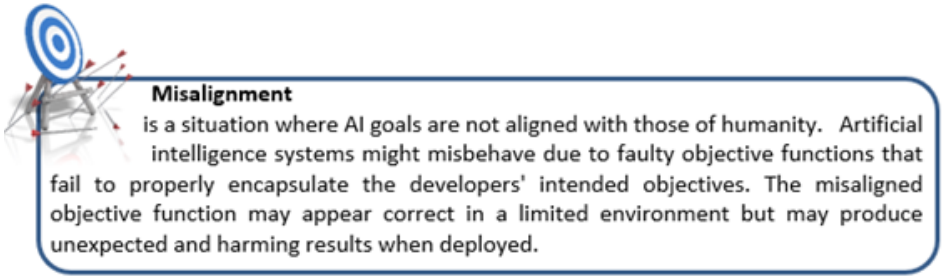



Figure IV.7 Misaligned goals in artificial intelligence (AI)

Current attempts to address the harmful effects of AI bias remain focused on computational factors. However, systemic, human, institutional, and societal factors are also important sources of AI bias and are currently overlooked. We hereby propose to initiate the discussion

and lay the groundwork for managing the risks associated with the use of AI in emergency medicine by identifying the biases that can be anticipated.

#### *IV.3.2.1. Bias in data and design*

Once end users (e.g. health care professionals) start interacting with an AI system or application, any early design and development decisions that were poorly specified and based on narrow perspectives can be exposed, leaving the process vulnerable to additive statistical or human biases[356].

##### Dataset bias challenge

Several categories of biases are held by health data sets used for training AI.
First, the choice of the data set for either pretraining or training can produce a sampling bias leading to a distributional shift[357], which is a mismatch between the data or environment in which the system is trained and that used in operation. Would training an AI application on EHRs of a local ED in a given region or state with given protocols and EHR architecture lead to the same results in the neighboring state's university hospital? When considering a physician-patient vocal assistant, how can language variety (regional or social dialects), linguistic variations (pronunciation, prosody, word choice, and grammar), and foreign speakers be considered?
Large-scale data sets are increasingly deployed for decision support applications, often in high-risk settings such as emergency medicine, and off-label uses result in representation bias harms. Low-represented populations or conditions should be carefully handled with rebalancing techniques such as data augmentation, oversampling, or weighting systems. Causal models and graphs can also be used to detect direct discrimination in the data [358,359].

Aggregation Bias (or ecological fallacy) arises when false conclusions are drawn about individuals from observing the entire population. An example of this type of bias in an emergency setting would be patients calling or presenting themselves with heart failure. Symptoms of heart failure differ in complex ways across genders[360,361]. Therefore, a model that ignores individual differences will likely not be well suited for gender groups in the population. This is true despite an equal representation in the training data. Any general assumptions regarding subgroups within the population can result in aggregation bias[362].

Simpson's paradox should also be considered at the designing step. The Simpson paradox is a type of aggregation bias that arises in the analysis of heterogeneous data[363]. The paradox arises when an association observed in aggregated data disappears or reverses when the same data are disaggregated into their underlying subgroups. For example, if an AI-guided CDSS was to be built for naloxone administration, when testing the model, if the clinical presentation severity or opioid type is unequally distributed among groups, the Simpson paradox will likely contribute to different rates of naloxone administration[364].

Modifiable Areal Unit Problem (MAUP) is a statistical bias in geospatial analysis that arises when modeling data at different levels of spatial aggregation[365]. This bias results in different trends learned when data are aggregated at different spatial scales. For example, when

designing an AI system for ambulance demand, only estimates based on minimal-resolution data should be relied upon, as ambulance demand using aerial data is potentially misleading due to the MAUP[366].

Omitted variable bias can also arise from variable selection for an emergency AI application. For example, when considering a triage application in which care protocols and treatment guidelines vary based on the patient's insurance status, omitting this variable could lead to errors in the triage score. However, considering this variable for better accuracy will lead to unfairness, which is already present in a real-world setting.

High quality input data are essential for the constructing a realistic AI system. Missing data bias is common in EHR data input quality management, and its gestion should be considered during the design step[367]. Several authors suggest that explicitly representing the presence or absence of data in the underlying logic of a CDSS can improve prediction performance[368]. Owing to the specificity of ED activities, data entry also comes with several biases such as recall bias (as health care practitioners often enter data several minutes or hours after the emergency has occurred), or confirmation bias (as healthcare practitioners often rely on heuristic-based decision[369]). It has recently been shown that serious games can improve physicians' heuristic judgment by providing them with a simulated experience. Additional experiments could lead to better data capture for less biased datasets[370].

Human biases, whether conditioned socially or cognitive, may influence data selection, preprocessing, annotation (attributing labels to an unlabeled dataset), and analysis process. Annotator biases could lead to biases in the training and test dataset. Hence, proper training on the annotation task, sufficient incentives, facilitating background and expertise diversity among annotators (e.g., nurses, physicians, researchers, and students), and the inclusion of a follow-up procedure with agreement evaluation could help in reducing these label biases[371].

Systemic institutional biases are also to be expected in the health datasets used to model underlying AI applications. The issue of "flattening" the societal and behavioral factors within the datasets themselves is problematic, but often overlooked[372]. If these biases are left unattended, AI applications are likely to reproduce human bias such as triage errors for women, the elderly and minor ethnicities[373,374].

### Bias in AI model choice and validation

The choice of models and their training process is a crucial step in the AI life cycle, and multiple biases can result from this. Most AI applications presented in the Actual and Possible Applications of AI for Emergency Services section are based on NLP, and concerns regarding the biases introduced by the growing use of large language models (ie, Bidirectional Encoder Representations from Transformers, Generative Pretrained Transformer 2, and XLNET) are relevant[375].

Semantic biases: Embeddings are the most common text inputs represented in NLP systems, and they have been shown to pick up on racial and gender biases in the training data[376]. As large language models are pretrained on almost the entire text corpus available from the

internet, they are prone to the same societal biases as those that prevail on the internet. Semantic biases hold not only for word embeddings but also for contextual representations. Debiasing sentence representation is at the heart of the efforts of some research teams. However, the impact and applicability of debiased embeddings are unclear for a wide range of downstream tasks[377].

Algorithmic effect: The algorithmic complexity can vary greatly from one AI model to another. The number of parameters that mathematically encode the training data can range from 1 to 1 trillion. Simple models with fewer parameters are often used because they tend to be cheaper to build, have better latency and better generalizability, are more explainable and transparent, and are easier to implement. However, these models can exacerbate statistical biases because restrictive assumptions about the training data often do not hold with nuanced demographic data. Complex models are often used for nonlinear and multimodal data such as text and images. These models can capture latent systemic biases in ways that are difficult to recognize and predict. Expert systems, another AI paradigm, can encode cognitive and perceptual biases in the accumulated knowledge of practitioners from which the system is designed to draw.

The objective function bias: The choice of the model's objective function, upon which the model's definition of accuracy is based, can reflect bias. In an emergency context, decisions must often be taken rapidly, meaning that AI should not increase the time required to reach a decision that would divert the patient to appropriate care. Not taking the vital and time context into consideration during model selection could harm patients. In addition to task-specific metrics, streaming and adaptation must be considered.

Validation bias: Performing tests on an AI system involved in health care under optimal conditions is challenging. Rigorous simulation and in-domain testing of time-specific windows or given locations should be performed before generalization. Randomized controlled trials and prospective studies in compliance with guidelines specific to AI interventions such as CONSORT-AI (Consolidated Standards of Reporting Trials–AI)[378] or SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials–AI)[379] should be conducted to ensure the transparency and validation of the application. The CONSORT-AI extension recommends that investigators provide clear descriptions of the AI intervention, including instructions and skills required for use, the setting in which the AI intervention is integrated, the handling of inputs and outputs of the AI intervention, the human-AI interaction, and the analysis of error cases.

#### *IV.3.2.2. Bias in deployment*

##### Inclusiveness bias

AI should encourage equitable use in emergency and primary care independent of age, gender, ethnicity, income, language spoken, or ability to comprehend. When considering a smartphone app or a digital lock at the entrance of an ED, different languages should be proposed. Accessibility devices for disabilities (visual, hearing, moving, and reading

impairments) should also be made available. Access to these technologies is particularly challenging for older adults, and alternative solutions should be proposed for this population.

#### Automation complacency

Health care practitioners may have a propensity to trust suggestions from AI decision support systems, which summarize large numbers of inputs into automated real-time predictions, while inadvertently discounting relevant information from nonautomated systems. Some information about the visual, behavioral, and intuitive analysis of a patient does not necessarily lead to rigorous documentation in EHR, yet this information contributes to clinical decision-making. Moreover, can this type of information can be captured by an AI model? Fully relying on a triage score prediction provided by an AI application without the necessary hindsight toward the added value of one's experience, common sense, and observation skills could lead to inaccurate resource allocation or priority levels for patients during triage.

#### Selective adherence

In contrast, health care practitioners can selectively adopt the AI advice when it matches their preexisting beliefs and stereotypes, leading to biases in the overall performance of the system.

### *IV.3.2.3. Monitoring*

Continuous measurement and monitoring of an algorithm's performance is necessary to assess whether it has a detrimental impact on patients or groups of patients. Tests and evaluations should cover the potential differential performance of the model according to age, gender, and relevant characteristics. As health care facilities benefit from quality and safety certification by public health and governmental agencies, AI technologies in health care should be audited periodically and externally. The report of these evaluations should be made public and intelligible to ensure transparency. In addition, assessing algorithm errors or deviations from human decisions can lead to reinforcement learning and an improvement in the model. Safe AI refers to the ability to modify misaligned systems. For this purpose, adversarial training procedures should be developed both as part of the training phase and the implementation.

### *IV.3.2.4. Fairness and inclusiveness*

#### Fairness

Fairness in AI includes concerns for equality and equity by addressing issues such as bias and discrimination. Fairness standards can be complex and difficult to define in emergency medicine because of disparities across health care systems (eg, in the United States, where hospital care protocols and treatment guidelines vary depending on the patient's insurance status), policies, and geographic areas.

### Inclusiveness

Inclusiveness requires that AI used in health care be tailored to support the broadest possible appropriate and equitable use and access, regardless of age, gender, income, ability, ethnicity, language spoken, or ability to comprehend. AI should be developed, deployed, and monitored by people from diverse disciplines, expertise, backgrounds, and cultures. AI technology should be designed and evaluated by those required to use the system including patients (who are themselves diverse).

## IV.3.3 Transparency, Accountability and Liability

In the interest of patient safety and trust, a certain amount of transparency must be ensured. Transparency reflects the extent to which information about an AI system or application is available to individuals. Its scope ranges from design decisions to training data, the structure of the model, its intended use case, and how and when deployment or end-user decisions were made and by whom. Transparency and participation can be increased by the use of open-source software for the underlying design of an AI technology or by making the source code of the software publicly available (eg, Babylon Health). However, there may be some legitime issues related to intellectual property protection[380].

The use of AI technologies in health care requires the assignment of responsibility within complex systems in which responsibility is distributed among different actors. When medical decisions made by AI technologies harm individuals, the responsibility and accountability processes must clearly identify the relative roles of manufacturers and clinical users in that harm. This is an evolving challenge that remains unsolved in the laws of most countries[381]. Institutions have not only a legal responsibility but also a duty to take responsibility for the decisions made by the algorithms they use. To avoid the diffusion of liability, a seamless liability model ("collective responsibility"), in which all stakeholders involved in the development and deployment of an AI technology are held accountable, can encourage all actors to act responsibly and minimize harm. Another proposition made by Maliha et al [382] is the creation of a compensation program that does not consider liability but instead assesses fees stakeholders.

Health care practitioners and health systems may be liable for malpractice or negligence. Imagine a dispatcher fully relying on an AI application that did not correctly classify the patient as high risk of having an OHCA, inducing delay in assistance and eventually death. To what extent would the dispatcher be liable for malpractice? So far, tort law protects health practitioners from liability as long as they follow the standards of care, regardless of its effectiveness in a particular case. AI involvement in emergency medicine has induced a previously unregulated paradigm shift. Possible legal outcomes depend on whether the AI application's recommendation follows the standard of care and on the AI accuracy, practitioner action, and patient outcome, as proposed by Price et al Table IV.1[383].

| AI recommendation | AI accuracy | Practitioner action | Patient outcome | Legal outcome (probable) |
|---|---|---|---|---|
| Standard of care | Correct | Follows | Good | No injury and no liability |
| | | Rejects | Bad | Injury and liability |
| | Incorrect (standard of care is incorrect) | Follows | Bad | Injury but no liability |
| | | Rejects | Good | No injury and no liability |
| Nonstandard of care | Correct (standard of care is incorrect) | Follows | Good | No injury and no liability |
| | | Rejects | Bad | Injury but no liability |
| | Incorrect | Follows | Bad | Injury and liability |
| | | Rejects | Good | No injury and no liability |

*Table IV.1 Examples of potential legal outcomes related to artificial intelligence (AI) use in clinical practice*[383]

Clinical malpractice, whether involving AI or not, leading to injury often induces compensation, as mentioned in Table IV.1 ED physicians already have higher rates of malpractice insurance owing to the higher risk of lawsuits. Does the malpractice insurer encompass the use of AI in high-risk fields such as emergency medicine? If so, how do we ensure that health care professionals receive the necessary insurance coverage? How can health care professionals be defended in court when they are threatened by claims involving AI? These questions remain to be answered by the legal community.

### IV.3.4 Explainability and Interpretability

Explainability refers to a representation of the mechanisms underlying the operation of an algorithm or model, whereas interpretability refers to the meaning of an AI system's output. Laws and regulations such as the European General Data Protection Regulation (GDPR) state that automated (or guided) decision-making should come along with the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject (Article 13{2}). When considering the possible application of emotion detection in voice during emergency calls to detect urgent conditions, the transparency and explainability of an AI solution is challenging. In emergency situations, the time requirements and explanation details collide. Thus, information regarding the outputs of an AI application should be meaningful and straightforward. Traditional machine learning models are mostly based on techniques that are inherently explainable. In contrast, deep learning models are considered as "black boxes" and have a higher computational cost (memory requirements and inference time). Explainable AI (XAI) is a recent field of research that attempts to provide solutions to confer trust in AI for practitioners[384]. XAI has additional features that enable better interpretability for end users. These features or explanations are provided for the model's process as a whole (global) or for an individual prediction (local). This explanation emerges directly from the prediction process (self-explaining) versus processing post hoc[385]. Depending on the stakeholder's expectations, the explanations and the way they are provided differ. There is a lack of consensus about which explanations can be used in different health care settings and how to measure them. Most studies have focused on subjective measurements, such as user satisfaction, goodness of explanation, acceptance, and trust in the system [386]. Further studies are required to evaluate the performance of XAI in health care settings.

### IV.3.5 Autonomy

#### *IV.3.5.1. For emergency health-care providers*

The adoption of AI in health care will lead to situations in which decision-making power can be, or is at least partially, transferred to machines. Protecting autonomy implies that humans remain in control of medical and health care system decisions. The opacity and "black-box" problem of an AI system[387] can make it difficult for health care professionals to ascertain how the system arrived at a decision and how an error may occur. How health care providers can be expected to remain in full control of their AI-assisted decisions when interpreting AI decisions is opaque even for developers. To what extent should health care providers inform patients that they do not fully interpret the recommendation provided by the AI system? AI systems should be designed to assist health care providers in making informed decisions. Moreover, to account for an AI application, ranking decisions and providing confidence score should be mandatory. For example, in the case of an emergency triage score, for each score proposed by an AI system, the predictions with highest accuracy should be given along with their associated probabilities.

#### *IV.3.5.2. For patients*

AI technology should not be used without the patient's valid informed consent. Owing to the patient's sometimes life-threatening condition, consent based on clear and intelligible information is not always feasible. Therefore, the responsibility for making an AI-assisted decision has shifted to health care professionals. Informed consent and its exceptions, without the use of AI, are equally regulated in the United States and Europe, with a tendency to not render practitioners liable for decisions taken in critical situations[388]. However, these statutory exceptions do not protect against litigation for malpractice and lack of informed consent[389]. Should health care practitioners use the AI-guided CDSS when obtaining informed consent is not possible? European Union has taken several steps to address the issue of liability when AI is involved in clinical decision-making. GDPR Article 13 (2): "[...] the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing: (f) the existence of automated decision-making, including profiling, referred to in Article 22 (1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."
Under Article 22 (1) and (3), "The data subject (i.e., the patient) shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her" unless the decision is "based on the data subject's explicit consent." However, the GDPR does not provide regulations for specific situations such as those mentioned Transparency, Accountability, and Liability section, but the European Commission is currently working on a liability directive to address and regulate liability for AI use [390,391].

### IV.3.6 Privacy-Enhanced

Privacy generally refers to norms and practices that help to preserve individual autonomy, identity, and dignity. Privacy-related values, such as anonymity, confidentiality, and control, should generally guide choices in the design, development, and deployment of AI systems. For example, the characteristics of AI and the novel risks associated with privacy protection are addressed in the European GDPR. Developing a compatible international framework to protect personal information would benefit stakeholders, and particularly patients, involved in AI for health care[392]. Clear information regarding the use of patient data for AI development purposes should be made available at any point of the emergency care trajectory. The right to erasure (right to be forgotten) as stated by GDPR Article 17 ("the data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay under given conditions") should be made possible, although it is problematic for AI developers.

## IV.4 Technical challenges

The implementation of systems using artificial intelligence or traditional machine learning in emergency services raises several issues on the practical and technical aspects.

### V.2.1 Training and data challenges

The quality, diversity and the size of the training dataset is a keystone for trustworthy AI applications. Acquiring or producing such datasets often is a time-consuming and expensive task. This procedure involves several entities collecting the data, transferring it to a central data repository, and fusing it to build a model. If EHR datasets originated from several hospitals were to be collected, the processes may violate laws such as the General Data Protection Regulation (GDPR) of the European Union, the California Consumer Privacy Act (CCPA), and Health Insurance Portability and Accountability Act (HIPAA)[393].

The setting up of a regulated and approved health data center or hub aimed at collecting all the necessary data is one of the future steps to be taken by the TARPON project. In compliance with laws and regulations, this type of center in conjunction with GPUs equipped servers will further push the boundaries of retrospective research. However, this type of center will not allow real-time data collection nor real-time modelling for a given patient and a given AI application.

Cloud-based digital health platforms could, on the other hand, bring together EHR, data connectivity, and powerful analytics. In doing so they address strategic issues for providers where monolithic EHR-centric application architectures cannot meet the changing demands of patients and clinical staff. A legal framework still has to be developed for this type of solution.

Recent work proposes the concept of federated learning (FL)[394] as seen on Figure IV.*8* to tackle privacy and technical issues. FL enables the training of AI models locally (at the location of the

data) and only shares the resulting model, which is not reverse-engineerable, with the requesting party. Therefore, FL avoids the need to share the private datasets and sensitive data to others, preventing exposition to entities conducting studies and enabling data usage for broader purposes[395]. A central entity manages the learning process and distributes the training algorithm to each participating data holder. Each participant generates a local model trained with their private data and shares the resulting parameters with the central entity. Finally, the central entity employs an aggregation algorithm to combine the parameters of all local models into a single global model.[396]

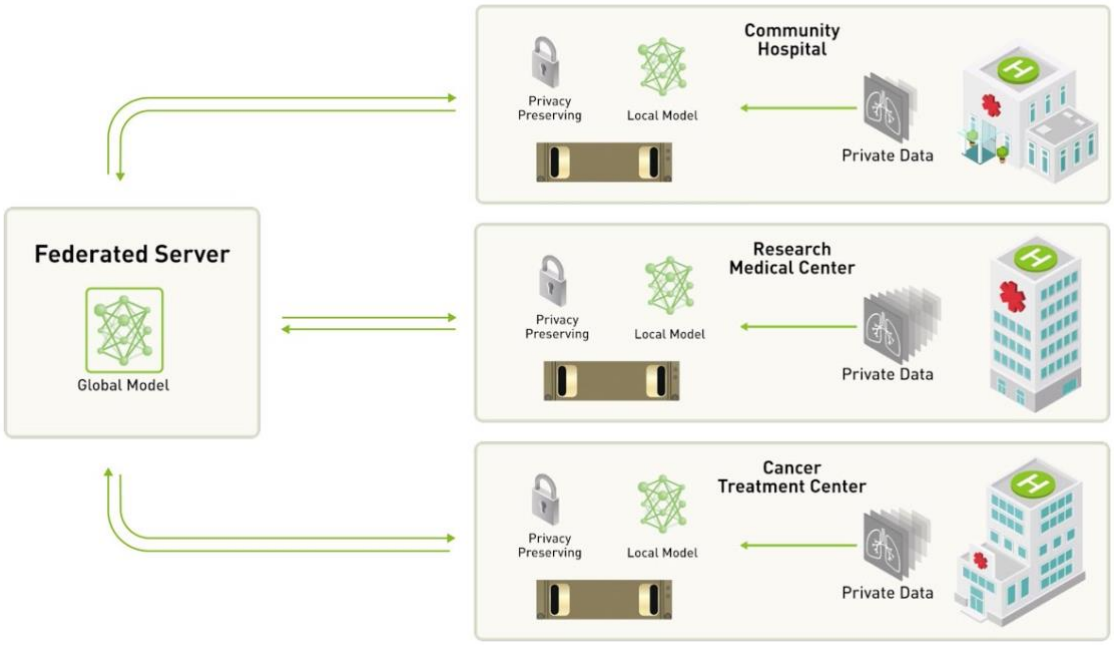



Figure IV.8 A centralized-server approach to federated learning (From [397])

### IV.4.1 Integration Into Routine Clinical Workflow

Once there is enough clinical evidence and a suitable regulatory framework in place, the integration of AI into the clinical workflow of an emergency department will encounter the final challenges in fully unlocking its potential.

The usability and ergonomics of AI applications is a key component of their integration into the clinical environment. Change resistance and full integration into the ED specific clinical environment should be considered before any implementation.

The second obstacle stems from the fact that numerous digital healthcare solutions have been developed as intricate monolithic systems. While these legacy systems are marketed as comprehensive digital healthcare solutions, their complexity and size make them challenging to use and integrate with current clinical systems. Additionally, they necessitate a substantial commitment from hospitals or healthcare systems interested in implementing the technology, as it often requires a complete overhaul of existing information technology systems. These factors combined contribute to significant hesitancy and hinder the adoption of artificial intelligence technologies[398].

To mitigate these challenges, AI programs could start as narrower, almost modular, solutions focused on specific clinical (for example, fast, accurate triaging in the ED). This would make them easier to integrate into existing clinical information technology infrastructure, allowing the technology to be adopted more quickly and across a broader range of health care environments[398].

## IV.5 Conclusion

AI has gained increasing attention owing to its potential advantages in health care and especially in emergency medicine for which several applications are currently used. Most ED and EMD AI applications are based on NLP and ASR because of the privileged documentation medium of free or semistructured text or the practitioner-patient interaction. There are limited studies on the types of models used and their validation methods. We noted a lack of evidence for symptom checkers with decreasing performance over time. Overall, AI-based applications in emergency medicine lack proper derivation, validation, or impact evaluations that are performed rigorously and independently.

Building a trustworthy, safe, and XAI requires a holistic approach that encompasses all sociotechnical aspects involved. Human factors such as participatory design and multistakeholder approaches are important for building such AI systems. Inclusiveness begins at the very beginning of the design step, with the inclusion of stakeholders (including end users) from diverse disciplines, expertise, backgrounds, and culture. All possible biases and risks should be identified and documented before any initiation, and they should be monitored continuously.

However, when emergency medicine is concerned with the development of AI applications, several principles mentioned above collide, and trade-offs must be determined. How can we determine the trade-off among interpretability and performance, time, and explainability? How can transparency be ensured when intellectual property is involved? How can liability be determined when AI harms?

AI should alleviate the high burden placed on health care professionals, but despite the ethical foundations laid, the actors gravitating around health care systems such as legislators, regulatory agencies, and insurers are not federated to ensure the safety of stakeholders.

# V. GENERAL CONCLUSION

Our contribution to the TARPON project laid the groundwork for the use of large language models for classifying clinical notes. These models, which are becoming increasingly efficient (accuracy) and powerful (with respect to database size), have led to a recent paradigm shift in NLP. It is likely that these models have not yet demonstrated their full potential in healthcare data (and elsewhere). However, unlike web-scraped corpora, health data corpora, especially in a language other than English, are difficult to build due to their level of sensitivity and legal protection. The weights of models trained on pseudonimized data are also an issue for sharing. Open-sourcing pre-trained models on EHR databases has little chance of success under current pre-training conditions. Investment in techniques for anonymizing any text from medical databases, with perfect performance and therefore 100% specificity, should be an absolute priority in the coming years. It will also be necessary to study the impact of data anonymization on the performance of classification, relation extraction and NER, since the legitimate question of information degradation arises. Another priority is to have sovereign models that do not depend on actors from outside the healthcare system and/or outside the country or the EU. Using the most efficient tools available today is equivalent to sending patient records to the United States.

One of the perspectives of the TARPON project are aimed at studying the impact of drugs on the risks of trauma, first at the local level in Bordeaux, then taking into account the 15 databases hosted by the Bordeaux hospital server. Ultimately, a fully functional system will generate signals on certain molecules and their associations thanks to a real time implementation of the data related to the trauma and a linkage with the SNDS. As the main purpose of TARPON is preventive, a prospective point of view seems to be the most efficient option. All the actors involved in the prevention of trauma and accidents in daily life share this vision. However, legal, technological and financial barriers need to be overcome. Meanwhile, technology related to artificial intelligence will continue to explode exponentially, making it even more difficult to align prevention to it. And in the face of resistance to change on the one hand, and regular paradigm shifts on the other, an area of overlap must be found in the interest of the people.

# VI. BIBLIOGRAPHY


1. Cour des Comptes. *6 Les Urgences Hospitalières : Des Services Toujours Trop Sollicités*.; 2019. Accessed March 2, 2020. https://www.ccomptes.fr/system/files/2019-02/08-urgences-hospitalieres-Tome-2.pdf
2. Baker SP, O'Neill B, Karpf RS. *The Injury Fact Book. Lexington, MA*. Lexington Books; 1984.
3. World Health Organization. International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)-WHO Version for ;2019 . Accessed March 2, 2020. https://icd.who.int/browse10/2019/en#/
4. World Health Organization. WHO Injuries. Accessed March 2, 2020. https://www.who.int/ceh/risks/cehinjuries2/en/
5. Holder Y, Peden M, Krug E. *WHO Injury Surveillance Guidelines*.; 2021. Accessed March 22, 2020. https://www.who.int/violence_injury_prevention/media/en/136.pdf
6. Kyu HH, Abate D, Abate KH, et al. Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*. 2018;392(10159):1859-1922.
7. The Global Burden of Disease (GBD) Team. Global Burden of Disease (GBD) Project. Accessed March 1, 2020. http://www.healthdata.org/gbd
8. Bège T, Pauly V, Orleans V, Boyer L, Leone M. Epidemiology of trauma in France: mortality and risk factors based on a national medico-administrative database. *Anaesth Crit Care Pain Med*. 2019;38(5):461-468.
9. World Health Organization. *WHO Guide to Identifying the Economic Consequences of Disease and Injury*.; 2009. Accessed February 2, 2020. https://www.who.int/choice/publications/d_economic_impact_guide.pdf
10. Krug EG, Sharma GK, Lozano R. The global burden of injuries. *Am J Public Health*. 2000;90(4):523.
11. Lyons RA, Turner S, Lyons J, et al. All Wales Injury Surveillance System revised: Development of a population-based system to evaluate single-level and multilevel interventions. *Injury Prevention*. 2016;22:i50-i55. doi:10.1136/injuryprev-2015-041814
12. Peden M and SR and SD and MD and HAA and JE and MCD and others. *World Report on Road Traffic Injury Prevention*.; 2004.
13. Observatoire national interministériel de la sécurité routière. *Bilan de l'accidentalité 2021*.; 2022. Accessed January 10, 2023. https://www.onisr.securite-routiere.gouv.fr/sites/default/files/2022-09/ONISR_Bilan_Accidentalit%C3%A9_2021_0.pdf
14. Observatoire National du Suicide. *Suicide: Enjeux éthiques de La Prévention, Singularités Suicide à l'adolescence*.; 2018.
15. World Health Organization. *Preventing Suicide: A Global Imperative*.; 2014. Accessed March 25, 2022. https://www.who.int/publications/i/item/9789241564779
16. Jardon V, Debien C, Duhem S, Morgiève M, Ducrocq F, Vaiva G. An example of post-discharge monitoring after a suicide attempt: VigilanS. *Encephale*. 2019;45:S13-S21. doi:10.1016/j.encep.2018.09.009

17. Plancke L, Amariei A, Danel T, et al. Effectiveness of a French Program to Prevent Suicide Reattempt (VigilanS). *Archives of Suicide Research*. 2021;25(3):570-581. doi:10.1080/13811118.2020.1735596
18. Amoros E, Martin JL, Laumon B. Under-reporting of road crash casualties in France. *Accid Anal Prev*. 2006;38(4):627-635.
19. Orriols L, Delorme B, Gadegbeku B, et al. Prescription medicines and the risk of road traffic crashes: A French registry-based study. *PLoS Med*. 2010;7(11). doi:10.1371/journal.pmed.1000366
20. Pfortmueller CA, Lindner G, Exadaktylos AK. Reducing fall risk in the elderly: risk factors and fall prevention, a systematic review. *Minerva Med*. 2014;105(4):275-281.
21. Bloch F, Thibaud M, Dugué B, Brèque C, Rigaud AS, Kemoun G. Psychotropic drugs and falls in the elderly people: Updated literature review and meta-analysis. *J Aging Health*. 2011;23(2):329-346. doi:10.1177/0898264310381277
22. Castro Madelyn Yiseth Rojas AND Orriols LANDCBANDDMANDSKCANDAMANDLE. Cohort profile: MAVIE a web-based prospective cohort study of home, leisure, and sports injuries in France. *PLoS One*. 2021;16(3):1-14. doi:10.1371/journal.pone.0248162
23. Moiron-Braud E, Générale De La Miprof S. *LES VIOLENCES CONJUGALES PENDANT LE CONFINEMENT : EVALUATION, SUIVI ET PROPOSITIONS*.; 2020.
24. State and Territorial Injury Prevention Directors Association and others. Safe States: Five components of a Model State Injury Prevention Program and Three Phases of Program Development. *Atlanta GA: STIPDA*. Published online 1997.
25. Mitchell RJ, Cameron CM, Bambach MR. Data linkage for injury surveillance and research in Australia perils. *Aust N Z J Public Health*. 2014;38(3):275-208.
26. Liu X, Li L, Cui H, Liu X, Jackson VW. Evaluation of an emergency department-based injury surveillance project in China using WHO guidelines. *Injury Prevention*. 2009;15(2):105. doi:10.1136/ip.2008.019877
27. Rogmans WHJ. Joint action on monitoring injuries in Europe (JAMIE). *Archives of Public Health*. 2012;70(1):19. doi:10.1186/0778-7367-70-19
28. Hémon D, Jougla E. *Surmortalité Liée à La Canicule d'août 2003 - Rapport d'étape Estimation de La Surmortalité et Principales Caractéristiques Épidémiologiques*.; 2003. Accessed February 10, 2022. https://www.inserm.fr/wp-content/uploads/2017-11/inserm-rapportthematique-surmortalitecaniculeaout2003-rapportetape.pdf
29. Buzyn A. *Communiqué de Presse, Agnès Buzyn, Ministre Des Solidarités et de La Santé Annonceles 10 Lauréats de l'appel à Projet Du Health Data Hub*.; 2019. https://solidarites-sante.gouv.fr/IMG/pdf/190412_dossier_de_presse_-
30. Health Data Hub. *HDH, Bilan 2022 & programme de Travail 2023*.; 2022. Accessed September 26, 2022. https://www.health-data-hub.fr/sites/default/files/2023-01/Bilan%202022%20et%20Programme%20de%20Travail%202023.pdf
31. Bourdois L, Avalos M, Chenais G, et al. De-identification of Emergency Medical Records in French: Survey and Comparison of State-of-the-Art Automated Systems. *Florida Artificial Intelligence Research Society*. 2021;34(1).
32. Agence du Numérique en Santé. Interopérabilité sémantique : la France choisit la SNOMED CT pour la description des localisations anatomiques. Published May 16, 2022. Accessed September 19, 2022. https://esante.gouv.fr/espace-presse/interoperabilite-

semantique-la-france-choisit-la-snomed-ct-pour-la-description-des-localisations-anatomiques
33. Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res*. 2004;32(DATABASE ISS.). doi:10.1093/nar/gkh061
34. Société Française de Médecine d'Urgence. Le thésaurus de Médecine d'Urgence. Published 2009. Accessed September 19, 2022. Le thésaurus de Médecine d'Urgence
35. World Health Organization. *International Classification of External Causes of Injuries (ICECI)*.; 2004. Accessed May 9, 2021. https://www.whofic.nl/sites/default/files/2018-05/ICECI%20in%20English.pdf
36. Smith CA, Hetzel S, Dalrymple P, Keselman A. Beyond readability: Investigating coherence of clinical text for consumers. *J Med Internet Res*. 2011;13(4). doi:10.2196/jmir.1842
37. Kvist bc M, Velupillai S, Wirén M. Improving Readability of Swedish Electronic Health Records through Lexical Simplification: First Results. In: *Association for Computational Linguistics*. ; :74-83. doi:10.3115/v1/W14-1209
38. Perry WM, Hossain R, Taylor RA. Assessment of the Feasibility of automated, real-time clinical decision support in the emergency department using electronic health record data. *BMC Emerg Med*. 2018;18(1). doi:10.1186/s12873-018-0170-9
39. Capurro D, Yetisgen M, Eaton E, Black R, Tarczy-Hornoch P. Availability of Structured and Unstructured Clinical Data for Comparative Effectiveness Research and Quality Improvement: A Multi-Site Assessment. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*. 2014;2(1):11. doi:10.13063/2327-9214.1079
40. Jo Y, Loghmanpour N, Rosé CP. Time Series Analysis of Nursing Notes for Mortality Prediction via a State Transition Topic Model. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. CIKM '15. Association for Computing Machinery; 2015:1171-1180. doi:10.1145/2806416.2806541
41. Joachims T. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*.
42. Gopinath D, Agrawal M, Murray L, Horng S, Karger D, Sontag D. Fast, Structured Clinical Documentation via Contextual Autocomplete Clinical Documentation with Contextual Autocomplete. *Proc Mach Learn Res*. 2020;106:1-26.
43. Smith CA, Hetzel S, Dalrymple P, Keselman A. Beyond readability: Investigating coherence of clinical text for consumers. *J Med Internet Res*. 2011;13(4). doi:10.2196/jmir.1842
44. Sørup FKH, Brunak S, Eriksson R. Association between antipsychotic drug dose and length of clinical notes: A proxy of disease severity? *BMC Med Res Methodol*. 2020;20(1). doi:10.1186/s12874-020-00993-1
45. Feldman K, Hazekamp N, Chawla N V. Mining the clinical narrative: all text are not equal. In: *2016 IEEE International Conference on Healthcare Informatics (ICHI)*. ; 2016:271-280.
46. Reed K, Doty DH, May DR. The Impact of Aging on Self-efficacy and Computer Skill Acquisition. *Journal of Managerial Issues*. 2005;17(2):212-228. http://www.jstor.org/stable/40604496
47. Dyck JL, Smither JAA. Age Differences in Computer Anxiety: The Role of Computer Experience, Gender and Education. *Journal of Educational Computing Research*. 1994;10(3):239-248. doi:10.2190/E79U-VCRC-EL4E-HRYV

48. Booth N, Robinson P, Kohannejad J. Identification of high-quality consultation practice in primary care: the effects of computer use on doctor–patient rapport. *Inform Prim Care*. 2004;12:75-83. doi:10.14236/jhi.v12i2.111
49. Frankel R, Altschuler A, George S, et al. Effects of exam-room computing on clinician-patient communication: A longitudinal qualitative study. *J Gen Intern Med*. 2005;20(8):677-682. doi:10.1111/j.1525-1497.2005.0163.x
50. Kadri F, Harrou F, Chaabane S, Tahon C. Time Series Modelling and Forecasting of Emergency Department Overcrowding. *J Med Syst*. 2014;38(9):107. doi:10.1007/s10916-014-0107-0
51. Newman ML, Groom CJ, Handelman LD, Pennebaker JW. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Process*. 2008;45(3):211-236. doi:10.1080/01638530802073712
52. Cunha E, Magno G, Gonçalves MA, Cambraia C, Almeida V. A Linguistic Characterization of Google+ Posts across Different Social Groups. In: *Notes of the 5th Workshop on Information in Networks (WIN)*. ; 2013. Accessed June 2, 2021. https://scholarlypublications.universiteitleiden.nl/access/item%3A2947519/view
53. Pennebaker J, King L. Linguistic styles: Language Use as an Individual Difference. *Journal of personality and social psychology*. 1999;77(6):1296. doi:10.1037/0022-3514.77.6.1296
54. Wang J, Deng H, Liu B, et al. Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: Bibliometric study on pubmed. *J Med Internet Res*. 2020;22(1). doi:10.2196/16816
55. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: Opportunities and challenges. *J Biomed Semantics*. 2018;9(1). doi:10.1186/s13326-018-0179-8
56. Wu S, Roberts K, Datta S, et al. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*. 2020;27(3):457-470. doi:10.1093/jamia/ocz200
57. Suominen H, Kelly L, Goeuriot L, et al. Overview of the CLEF ehealth evaluation lab 2018. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 11018 LNCS. Springer Verlag; 2018:286-301. doi:10.1007/978-3-319-98932-7_26
58. Cossin S, Jouhet V, Mougin F, Diallo G, Thiessard F. IAM at CLEF eHealth 2018: Concept Annotation and Coding in French Death Certificates. Published online July 10, 2018. http://arxiv.org/abs/1807.03674
59. Flicoteaux R. ECSTRA-APHP @ CLEF eHealth2018-task 1: ICD10 Code Extraction from Death Certificates. In: *In CLEF (Working Notes).* ; 2018. http://www.cepidc.inserm.fr/
60. Amin-Nejad A, Ive J, Velupillai S. *Exploring Transformer Text Generation for Medical Dataset Augmentation*.; 2020. https://github.com/tensorflow/tensor2tensor
61. CHU Rouen. HeTOP (Health Terminology/Ontology Portal). Accessed September 19, 2021. https://www.hetop.eu/hetop/rep/terminologies/
62. Metzger MH, Tvardik N, Gicquel Q, Bouvry C, Poulet E, Potinet-Pagliaroli V. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a French pilot study. *Int J Methods Psychiatr Res*. 2017;26(2). doi:10.1002/mpr.1522

63. Gerbier S, Yarovaya O, Gicquel Q, et al. Evaluation of natural language processing from emergency department computerized medical records for intra-hospital syndromic surveillance. *BMC Med Inform Decis Mak*. 2011;11(1). doi:10.1186/1472-6947-11-50
64. Soualmia LF, Dahamna B, Thirion B, Darmoni SJ. *Strategies for Health Information Retrieval*.; 2006. http://www.chu-rouen.fr/cismef.
65. Li I, Pan J, Goldwasser J, et al. Neural Natural Language Processing for unstructured data in electronic health records: A review. *Comput Sci Rev*. 2022;46:100511. doi:https://doi.org/10.1016/j.cosrev.2022.100511
66. Vaswani A, Brain G, Shazeer N, et al. Attention Is All You Need. In: *Advances in Neural Information Processing Systems*. ; 2017. doi:10.48550/arXiv.1706.03762
67. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Published online October 10, 2018. http://arxiv.org/abs/1810.04805
68. Le H, Vial L, Frej J, et al. FlauBERT: Unsupervised Language Model Pre-training for French. Published online December 11, 2019. http://arxiv.org/abs/1912.05372
69. Martin L, Muller B, Suárez PJO, et al. CamemBERT: a Tasty French Language Model. Published online November 10, 2019. doi:10.18653/v1/2020.acl-main.645
70. Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Published online July 26, 2019. http://arxiv.org/abs/1907.11692
71. Javier Ortiz Suárez P, Sagot B, Romary L, Sagot B. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. Published online 2019. doi:10.14618/IDS-PUB
72. Wenzek G, Lachaux MA, Conneau A, et al. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. Published online November 1, 2019. http://arxiv.org/abs/1911.00359
73. Louis A. BelGPT-2: a GPT-2 model pre-trained on French corpora. Published 2020. https://github.com/antoiloui/belgpt2
74. Lopez-Garcia G, Jerez JM, Ribelles N, Alba E, Veredas FJ. Transformers for Clinical Coding in Spanish. *IEEE Access*. 2021;9:72387-72397. doi:10.1109/ACCESS.2021.3080085
75. Zhang Z, Liu J, Razavian N. BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining. Published online May 26, 2020. http://arxiv.org/abs/2006.03685
76. Liu Z, He H, Yan S, Wang Y, Yang T, Li GZ. End-to-end models to imitate traditional chinese medicine syndrome differentiation in lung cancer diagnosis: Model development and validation. *JMIR Med Inform*. 2020;8(6). doi:10.2196/17821
77. Mohammadi R, Jain S, Namin AT, et al. Predicting unplanned readmissions following a hip or knee arthroplasty: retrospective observational study. *JMIR Med Inform*. 2020;8(11). doi:10.2196/19761
78. Witten IH, Frank E, Hall MA, Pal CJ, DATA M. *Practical Machine Learning Tools and Techniques*. Vol 2. Elsevier; 2005. doi:10.1016/c2009-0-19715-5
79. Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. *AI Mag*. 1996;17. doi:10.1609/aimag.v17i3.1230
80. Sharmila K, Vethamanickam SA. Survey on Data Mining Algorithm and Its Application in Healthcare Sector Using Hadoop Platform. *International Journal of Emerging Technology and Advanced Engineering*. 2008;9001(1).

81. Kharya S. Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease. *International Journal of Computer Science, Engineering and Information Technology*. 2012;2(2):55-66. doi:10.5121/ijcseit.2012.2206
82. Lashari SA, Ibrahim R, Senan N. De-noising Analysis of Mammogram Images in the Wavelet Domain using Hard and Soft Thresholding. In: *2014 4th World Congress on Information and Communication Technologies (WICT 2014)*. ; 2014:353-357. doi:10.1109/wict.2014.7077293
83. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform*. 2009;42(5):760-772. doi:10.1016/j.jbi.2009.08.007
84. Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics*. 2013;14(1):10. doi:10.1186/1471-2105-14-10
85. Liu' H, Lussier' YA, Friedman2' C. A Study of Abbreviations in the UMLS. In: *Proc AMIA Symp.* ; 2001.
86. Quimbaya AP, Múnera AS, Rivera RAG, et al. Named Entity Recognition over Electronic Health Records Through a Combined Dictionary-based Approach. In: *Procedia Computer Science*. Vol 100. Elsevier B.V.; 2016:55-61. doi:10.1016/j.procs.2016.09.123
87. Pitoglou S, Filntisi A, Anastasiou A, Matsopoulos GK, Koutsouris D. Exploring the Utility of Anonymized EHR Datasets in Machine Learning Experiments in the Context of the MODELHealth Project. *Applied Sciences (Switzerland)*. 2022;12(12). doi:10.3390/app12125942
88. Iglewicz B, Hoaglin DC. *Volume 16: How to Detect and Handle Outliers*. Quality Press; 1993.
89. Apathy NC, Hare AJ, Fendrich S, Cross DA. I had not time to make it shorter: an exploratory analysis of how physicians reduce note length and time in notes. *Journal of the American Medical Informatics Association*. 2023;30(2):355-360. doi:10.1093/jamia/ocac211
90. Huang AE, Hribar MR, Goldstein IH, Henriksen B, Lin WC, Chiang MF. *Clinical Documentation in Electronic Health Record Systems: Analysis of Similarity in Progress Notes from Consecutive Outpatient Ophthalmology Encounters*.
91. Zhelezniak V, Savkov A, Shen A, Hammerla N. Correlation Coefficients and Semantic Textual Similarity. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics; 2019:951-962. doi:10.18653/v1/N19-1100
92. Blei DM, Ng AY, Edu JB. *Latent Dirichlet Allocation Michael I. Jordan*. Vol 3.; 2003.
93. Zuo Y, Zhao J, Xu K. Word Network Topic Model: A Simple but General Solution for Short and Imbalanced Texts. *Knowl Inf Syst*. 2016;48(2):379-398. doi:10.1007/s10115-015-0882-z
94. Zhao R, Mao K. Supervised Adaptive-Transfer PLSA for Cross-Domain Text Classification. In: *2014 IEEE International Conference on Data Mining Workshop*. ; 2014:259-266. doi:10.1109/ICDMW.2014.163
95. Yan X, Guo J, Lan Y, Cheng X. A Biterm Topic Model for Short Texts. In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW '13. Association for Computing Machinery; 2013:1445-1456. doi:10.1145/2488388.2488514

96. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. Published online March 11, 2022. http://arxiv.org/abs/2203.05794
97. McInnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering. *J Open Source Softw*. 2017;2(11):205.
98. de Groot M, Aliannejadi M, Haas MR. Experiments on Generalizability of BERTopic on Multi-Domain Short Text. Published online December 16, 2022. http://arxiv.org/abs/2212.08459
99. Lebeña N, Blanco A, Pérez A, Casillas A. Preliminary exploration of topic modelling representations for Electronic Health Records coding according to the International Classification of Diseases in Spanish. *Expert Syst Appl*. 2022;204. doi:10.1016/j.eswa.2022.117303
100. Goldberg Y. *Neural Network Methods for Natural Language Processing*. Vol 10. Springer International Publishing; 2017. doi:10.1007/978-3-031-02165-7_5
101. Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*. 2004;60(5):503-520. doi:10.1108/00220410410560582
102. Fong ACM. A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*. 2010;1(1). doi:10.4304/jait.1.1.1-1
103. Jacobson O, Dalianis H. *Applying Deep Learning on Electronic Health Records in Swedish to Predict Healthcare-Associated Infections*.; 2016. http://snowball.tartarus.org/algorithms/swedish/stemmer.html.
104. Sieper AA, Amarkhel O, Diez S, Petrak D. *Semantic Code Search with Neural Bag-of-Words and Graph Convolutional Networks*.; 2020. https://github.blog/2019-09-26-introducing-the-codesearchnet-challenge/
105. Church KW. Emerging Trends: Word2Vec. *Nat Lang Eng*. 2017;23(1):155-162. doi:10.1017/S1351324916000334
106. Pennington J, Socher R, Manning CD. *GloVe: Global Vectors for Word Representation*.; 2014. http://nlp.
107. Gage P. A New Algorithm for Data Compression. *C Users Journal*. 1994;12(2):23-38. doi:10.5555/177910.177914
108. Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units. *ArXiv*. Published online August 31, 2015. http://arxiv.org/abs/1508.07909
109. Yang S, Dengyong Z, Sanqiang Z. *Extreme Language Model Compression with Optimal Subwords and Shared Projections*.; 2021. Accessed June 2, 2021. https://patents.google.com/patent/US20210224660A1/en
110. Tang R, Lu Y, Liu L, Mou L, Vechtomova O, Lin J. Distilling Task-Specific Knowledge from BERT into Simple Neural Networks. *ArXiv*. Published online March 28, 2019. http://arxiv.org/abs/1903.12136
111. Peters ME, Neumann M, Iyyer M, et al. Deep contextualized word representations. *ArXiv*. Published online February 14, 2018. http://arxiv.org/abs/1802.05365
112. Wang C, Cho K, Gu J. Neural Machine Translation with Byte-Level Subwords. In: *AAAI Conference on Artificial Intelligence*. ; 2020:9154-9160. www.aaai.org
113. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*. Published online October 2, 2019. http://arxiv.org/abs/1910.01108

114. Clark K, Luong MT, Le Q V., Manning CD. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *ArXiv*. Published online March 23, 2020. http://arxiv.org/abs/2003.10555
115. Institute of Electrical and Electronics Engineers., IEEE Signal Processing Society. *2012 IEEE International Conference on Acoustics, Speech, and Signal Processing : Proceedings : March 25-30, [2012], Prague Congress Center, Prague, Czech Republic*. IEEE; 2012.
116. Abraham R, Simha JB, Iyengar SS. Medical Datamining with a New Algorithm for Feature Selection and Naive Bayesian Classifier. In: *10th International Conference on Information Technology*. Institute of Electrical and Electronics Engineers (IEEE); 2008:44-49. doi:10.1109/icit.2007.41
117. Ehsani-Moghaddam B, Queenan JA, MacKenzie J, Birtwhistle R V. Mucopolysaccharidosis type II detection by Na\"\ive Bayes Classifier: An example of patient classification for a rare disease using electronic medical records from the Canadian Primary Care Sentinel Surveillance Network. *PLoS One*. 2018;13(12):e0209018.
118. Duda RO, Hart PE, others. *Pattern Classification and Scene Analysis*. Vol 3. Wiley New York; 1973.
119. Breiman L. *Classification and Regression Trees*. Routledge; 2017.
120. Quinlan JR. *C4. 5: Programs for Machine Learning*. Elsevier; 2014.
121. Kass G V. An exploratory technique for investigating large quantities of categorical data. *J R Stat Soc Ser C Appl Stat*. 1980;29(2):119-127. doi:10.2307/2986296
122. Breiman L. Bagging predictors. *Mach Learn*. 1996;24:123-140. doi:10.1007/bf00058655
123. Freund Y, Schapire RE, others. Experiments with a new boosting algorithm. In: *Icml*. Vol 96. ; 1996:148-156.
124. Schapire RE, Singer Y. BoosTexter: A boosting-based system for text categorization. *Mach Learn*. 2000;39:135-168.
125. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: A literature review. *J Biomed Inform*. 2018;77:34-49. doi:10.1016/j.jbi.2017.11.011
126. Houssein EH, Mohamed RE, Ali AA. Machine Learning Techniques for Biomedical Natural Language Processing: A Comprehensive Review. *IEEE Access*. 2021;9:140628-140653. doi:10.1109/ACCESS.2021.3119621
127. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J Biomed Inform*. 2015;53:196-207. doi:10.1016/j.jbi.2014.11.002
128. Roberts K, Rink B, Harabagiu SM, et al. A Machine Learning Approach for Identifying Anatomical Locations of Actionable Findings in Radiology Reports. In: *AMIA Annu Symp Proc*. ; 2012:779-788.
129. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444.
130. Mitchell TM, others. *Machine Learning*. Vol 1. McGraw-hill New York; 2007.
131. Deng L, Yu D, others. Deep learning: methods and applications. *Foundations and trends® in signal processing*. 2014;7(3–4):197-387.
132. CNN image. Accessed April 27, 2022. https://www.mathworks.com/videos/introduction-to-deep-learning-what-are-convolutional-neural-networks--1489512765771.html

133. Hughes M, Li I, Kotoulas S, Suzumura T. Medical Text Classification Using Convolutional Neural Networks. *Stud Health Technol Inform*. 2017;235:246-250. doi:10.3233/978-1-61499-753-5-246
134. Wang Y, Sohn S, Liu S, et al. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak*. 2019;19(1):1. doi:10.1186/s12911-018-0723-6
135. Olah C. Understanding LSTM Networks. Accessed April 26, 2021. https://colah.github.io/posts/2015-08-Understanding-LSTMs/
136. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. 1998;6(02):107-116.
137. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-1780.
138. Cho K, van Merriënboer B, Bahdanau D, Bengio Y. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. ; 2014:103-111.
139. Futoma J, Hariharan S, Heller K. Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier. In: Precup D, Teh YW, eds. *Proceedings of the 34th International Conference on Machine Learning*. Vol 70. Proceedings of Machine Learning Research. PMLR; 2017:1174-1182. https://proceedings.mlr.press/v70/futoma17a.html
140. Yan RQ, Liu W, Zhu M, et al. Real-time abnormal light curve detection based on a Gated Recurrent Unit network. *Res Astron Astrophys*. 2020;20(1):7.
141. Zhou C, Li Q, Li C, et al. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. *ArXiv*. Published online 2023.
142. Alammar J. The Illustrated Transformer. . Published 2018. Accessed May 17, 2020. https://jalammar.github.io/illustrated-transformer/
143. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:210807258*. Published online 2021.
144. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le Q V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *ArXiv*. Published online June 19, 2019. http://arxiv.org/abs/1906.08237
145. McCandlish S, Kaplan J, Amodei D, Team OD. An empirical model of large-batch training. *arXiv preprint arXiv:181206162*. Published online 2018.
146. Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. *arXiv preprint arXiv:200108361*. Published online 2020.
147. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:14126980*. Published online 2014.
148. Zhang J, Karimireddy SP, Veit A, et al. Why adam beats sgd for attention models. Published online 2019.
149. Erhan D, Bengio Y, Courville A, Manzagol PA, Vincent P, Bengio S. Why Does Unsupervised Pre-training Help Deep Learning? *Journal of Machine Learning Research*. 2010;11(19):625-660. http://jmlr.org/papers/v11/erhan10a.html
150. Wang Y, Liu S, Afzal N, et al. A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform*. 2018;87:12-20. doi:10.1016/j.jbi.2018.09.008

151. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. Published online April 10, 2019. http://arxiv.org/abs/1904.05342
152. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3(1):160035. doi:10.1038/sdata.2016.35
153. Alsentzer E, Murphy J, Boag W, et al. Publicly Available Clinical BERT Embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2019:72-78. doi:10.18653/v1/W19-1909
154. Dieng AB, Ruiz FJR, Blei DM. Topic Modeling in Embedding Spaces. *Trans Assoc Comput Linguist*. 2020;8:439-453. doi:10.1162/tacl_a_00325
155. Campillos L, Deléger L, Grouin C, Hamon T, Ligozat AL, Névéol A. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). *Lang Resour Eval*. 2018;52(2):571-601. doi:10.1007/s10579-017-9382-y
156. Grabar N, Dalloux C, Claveau V. CAS: Corpus of clinical cases in French. *J Biomed Semantics*. 2020;11(1). doi:10.1186/s13326-020-00225-x
157. Hiebel N, Ferret O, Fort K, Névéol A. *CLISTER: A Corpus for Semantic Textual Similarity in French Clinical Narratives*.; 2022. https://deft.limsi.fr/2020/
158. Lee J, Yoon W, Kim S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240. doi:10.1093/bioinformatics/btz682
159. Conneau A, Lample G. Cross-lingual Language Model Pretraining. In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. ; 2019. https://github.com/facebookresearch/XLM
160. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *ArXiv*. Published online September 26, 2019. http://arxiv.org/abs/1909.11942
161. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*. 2020;21(1):5485-5551.
162. Chi Z, Huang S, Dong L, et al. XLM-E: Cross-lingual Language Model Pre-training via ELECTRA. *ArXiv*. Published online June 30, 2021. http://arxiv.org/abs/2106.16138
163. Soltan S, Ananthakrishnan S, FitzGerald J, et al. AlexaTM 20B: Few-Shot Learning Using a Large-Scale Multilingual Seq2Seq Model. *ArXiv*. Published online August 2, 2022. http://arxiv.org/abs/2208.01448
164. Zoph B, Bello I, Kumar S, et al. ST-MoE: Designing Stable and Transferable Sparse Expert Models. *ArXiv*. Published online February 17, 2022. http://arxiv.org/abs/2202.08906
165. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*. Vol 33. ; 2020:1877-1901.
166. Zhang S, Roller S, Goyal N, et al. OPT: Open Pre-trained Transformer Language Models. *ArXiv*. Published online May 2, 2022. http://arxiv.org/abs/2205.01068
167. Chowdhery A, Narang S, Devlin J, et al. PaLM: Scaling Language Modeling with Pathways. *ArXiv*. Published online April 5, 2022. http://arxiv.org/abs/2204.02311
168. Workshop B, :, Scao T Le, et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *ArXiv*. Published online November 9, 2022. http://arxiv.org/abs/2211.05100

169. Smith S, Patwary M, Norick B, et al. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. *ArXiv*. Published online January 28, 2022. http://arxiv.org/abs/2201.11990
170. Du N, Huang Y, Dai AM, et al. Glam: Efficient scaling of language models with mixture-of-experts. In: *International Conference on Machine Learning*. ; 2022:5547-5569.
171. Rae JW, Borgeaud S, Cai T, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:211211446*. Published online 2021.
172. Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models. *arXiv preprint arXiv:220315556*. Published online 2022.
173. Thoppilan R, De Freitas D, Hall J, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:220108239*. Published online 2022.
174. Wang B. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. Published online May 2021.
175. Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:230213971*. Published online 2023.
176. Wu S, Irsoy O, Lu S, et al. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:230317564*. Published online 2023.
177. Driess D, Xia F, Sajjadi MSM, et al. PaLM-E: An Embodied Multimodal Language Model. *ArXiv*. Published online March 6, 2023. http://arxiv.org/abs/2303.03378
178. Yang J, Jin H, Tang R, et al. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ArXiv*. Published online 2023.
179. Hugging Face Blog. Accessed May 17, 2022. https://huggingface.co/blog
180. Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:160908144*. Published online 2016.
181. Zhu Y, Kiros R, Zemel R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: *Proceedings of the IEEE International Conference on Computer Vision*. ; 2015:19-27.
182. Dodge J, Ilharco G, Schwartz R, Farhadi A, Hajishirzi H, Smith N. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. *ArXiv*. Published online 2020.
183. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med*. 2021;4(1). doi:10.1038/s41746-021-00455-y
184. Li Y, Rao S, Solares JRA, et al. BEHRT: Transformer for Electronic Health Records. *Sci Rep*. 2020;10(1):7155. doi:10.1038/s41598-020-62922-y
185. Shang J, Ma T, Xiao C, Sun J. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:190600346*. Published online 2019.
186. Labrak Y, Bazoge A, Dufour R, et al. DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains. *ArXiv*. Published online 2023.
187. Valmianski I, Sood N, Wang YangX Y, et al. *SmartTriage: A System for Personalized Patient Data Capture, Documentation Generation, and Decision Support*. Vol 158.; 2021.
188. Radford A, Narasimhan K, Salimans T, Sutskever I, others. Improving language understanding by generative pre-training. In: OpenAI; 2018.

189. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. *OpenAI blog*. 2019;1(8)(9). https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
190. Liu PJ, Saleh M, Pot E, et al. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:180110198*. Published online 2018.
191. Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. *Adv Neural Inf Process Syst*. 2017;30.
192. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Published online 2022.
193. Rahman MM, Davis DN. Addressing the Class Imbalance Problem in Medical Datasets. *Int J Mach Learn Comput*. Published online 2013:224-228. doi:10.7763/ijmlc.2013.v3.307
194. Khushi M, Shaukat K, Alam TM, et al. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access*. 2021;9:109960-109975. doi:10.1109/ACCESS.2021.3102399
195. Ishwaran H, O'Brien R. Commentary: The problem of class imbalance in biomedical data. *Journal of Thoracic and Cardiovascular Surgery*. 2021;161(6):1940-1941. doi:10.1016/j.jtcvs.2020.06.052
196. Coulombe C. Text Data Augmentation Made Simple By Leveraging NLP Cloud APIs. Published online 2018.
197. Miller GA. WordNet: a lexical database for English. *Commun ACM*. 1995;38(11):39-41.
198. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. Published online 2017.
199. Wei J, Zou K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. Published online 2019:6382-6388. http://github.
200. Wang WY, Yang D. *That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors Using #petpeeve Tweets*. Association for Computational Linguistics; 2015. http://www.cs.cmu.edu/
201. Zhang Y, Ge T, Sun X. Parallel data augmentation for formality style transfer. *arXiv preprint arXiv:200507522*. Published online 2020.
202. Anaby-Tavor A, Carmeli B, Goldbraich E, et al. Do not have enough data? Deep learning to the rescue! In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol 34. ; 2020:7383-7390.
203. Ng N, Cho K, Ghassemi M. SSMBA: Self-Supervised Manifold Based Data Augmentation for Improving Out-of-Domain Robustness. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ; 2020:1268-1283.
204. Zhang D, Li T, Zhang H, Yin B. On data augmentation for extreme multi-label classification. *arXiv preprint arXiv:200910778*. Published online 2020.
205. Quteineh H, Samothrakis S, Sutcliffe R. Textual data augmentation for efficient active learning on tiny datasets. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ; 2020:7400-7410.
206. Tarján B, Szaszák G, Fegyó T, Mihajlik P. Deep Transformer based Data Augmentation with Subword Units for Morphologically Rich Online ASR. *arXiv preprint arXiv:200706949*. Published online 2020.

207. Pankaj S, Gautam A. Augmented Bio-SBERT: Improving Performance for Pairwise Sentence Tasks in Bio-medical Domain. In: *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*. ; 2022:43-47.
208. Miao L, Last M, Litvak M. Twitter data augmentation for monitoring public opinion on COVID-19 intervention measures. In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. ; 2020.
209. Montella S, Fabre B, Urvoy T, Heinecke J, Barahona LMR. Denoising Pre-Training and Data Augmentation Strategies for Enhanced RDF Verbalization with Transformers. In: *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*. ; 2020:89-99.
210. Perevalov A, Both A. Augmentation-based Answer Type Classification of the SMART dataset. Published online 2020.
211. Grandini M, Bagli E, Visani G. Metrics for Multi-Class Classification: an Overview. Published online August 13, 2020. http://arxiv.org/abs/2008.05756
212. Arzucan¨ozgür, Levent¨ozgür L, Güngör T. Text Categorization with Class-Based and Corpus-Based Keyword Selection. In: *Computer and Information Sciences (ISCIS 2005)*. ; 2005:606-615.
213. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(1). doi:10.1186/s12864-019-6413-7
214. Rehurek R, Sojka P. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*. 2011;3(2):2.
215. Jurafsky D, Martin JH. Speech and language processing (draft). *Chapter A: Hidden Markov Models (Draft of September 11, 2018) Retrieved March*. 2018;19:2019.
216. french-camembert-postag-model. Accessed May 3, 2022. https://huggingface.co/gilf/french-camembert-postag-modeluggingface.co/gilf/french-camembert-postag-model
217. Hernandez N. French Tree Bank. Accessed May 3, 2022. https://github.com/nicolashernandez/free-french-treebank
218. Abeillé A, Clément L, Toussenel F. Building a treebank for French. *Treebanks: Building and using parsed corpora*. Published online 2003:165-187.
219. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:190810084*. Published online 2019.
220. Reimers N, Gurevych I. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:200409813*. Published online 2020.
221. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426*. Published online 2018.
222. Landis JR, Koch GG. *The Measurement of Observer Agreement for Categorical Data*. Vol 33.; 1977.
223. Bird S. NLTK: The Natural Language Toolkit. In: *Proceediproceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. ; 2006:69-72. https://aclanthology.org/P06-4018.pdf
224. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825-2830.

225. Edgar TW, Manz DO. Chapter 4 - Exploratory Study. In: Edgar TW, Manz DO, eds. *Research Methods for Cyber Security*. Syngress; 2017:95-130. doi:https://doi.org/10.1016/B978-0-12-805349-2.00004-2
226. Platt J, others. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*. 1999;10(3):61-74.
227. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*. 2011;2(3):1-27.
228. Tsuruoka Y, Tsujii J, Ananiadou S. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. ; 2009:477-485.
229. Xu W. Towards optimal one pass large scale learning with averaged stochastic gradient descent. *arXiv preprint arXiv:11072490*. Published online 2011.
230. Abdi H, Valentin D, Edelman B, O'Toole AJ. More about the difference between men and women: evidence from linear neural networks and the principal-component approach. *Perception*. 1995;24(5):539-562.
231. Crammer K, Dekel O, Keshet J, Shalev-Shwartz S, Singer Y. Online passive aggressive algorithms. Published online 2006.
232. Rifkin RM, Lippert RA. Notes on regularized least squares. Published online 2007.
233. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn*. 2006;63:3-42.
234. Rennie JD, Shih L, Teevan J, Karger DR. Tackling the poor assumptions of naive bayes text classifiers. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. ; 2003:616-623.
235. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30.
236. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol 13-17-August-2016. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785
237. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30.
238. Ortiz Suárez PJ, Sagot B, Romary L. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In: Bański P, Barbaresi A, Biber H, et al., eds. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019. Leibniz-Institut für Deutsche Sprache; 2019:9 – 16. doi:10.14618/ids-pub-9021
239. Komatsuzaki A. One Epoch Is All You Need. Published online June 16, 2019. http://arxiv.org/abs/1906.06669
240. Hernandez N. french-camembert-postag-model. Accessed May 4, 2021. https://huggingface.co/gilf/french-camembert-postag-model
241. Hernandez N, Boudin F. Construction automatique d'un large corpus libre annot'e morpho-syntaxiquement en français. In: *Actes de la conf'erence TALN-RECITAL 2013*. ; 2013.
242. Chenais G, Gil-Jardiné C, Touchais H, et al. Deep Learning Transformer Models for Building a Comprehensive and Real-time Trauma Observatory: Development and Validation Study. *JMIR AI*. 2023;2:e40843. doi:10.2196/40843

243. Chenais Gabrielle and Gil-Jardiné C and TH and AFM and CB and TE and CX and BL and RP and LE. Deep Learning Transformer Models for Building a Comprehensive and Real-time Trauma Observatory: Development and Validation Study. *JMIR AI*. 2023;2:e40843. doi:10.2196/40843
244. Gil-Jardiné C, Chenais G, Pradeau C, et al. Trends in reasons for emergency calls during the COVID-19 crisis in the department of Gironde, France using artificial neural network for natural language classification. *Scand J Trauma Resusc Emerg Med*. 2021;29(1). doi:10.1186/s13049-021-00862-w
245. World Health Organization. *International Statistical Classification of Diseases and Related Health Problems: Tenth Revision (ICD-10). 10th Revision*.; 1992.
246. World Health Organization. *"World Health Assembly Update, 25 May 2019" (Press Release)*.; 2019. Accessed February 3, 2023. https://www.who.int/news-room/detail/25-05-2019-world-health-assembly-update
247. FEDORU. *Le RPU V3 : Propositions de La FEDORU*.; 2020. Accessed April 5, 2022. https://fedoru.fr/fiche-publications/le-rpu-v3-propositions-de-la-fedoru/
248. Pines JM, Pollack C V., Diercks DB, Chang AM, Shofer FS, Hollander JE. The association between emergency department crowding and adverse cardiovascular outcomes in patients with chest pain. *Academic Emergency Medicine*. 2009;16(7):617-625. doi:10.1111/j.1553-2712.2009.00456.x
249. Richardson DB. Increase in patient mortality at 10 days associated with emergency department overcrowding. *Medical Journal of Australia*. 2006;184(5):213-216. doi:10.5694/j.1326-5377.2006.tb00204.x
250. Hooker EA, Mallow PJ, Oglesby MM. Characteristics and Trends of Emergency Department Visits in the United States (2010–2014). *Journal of Emergency Medicine*. 2019;56(3):344-351. doi:10.1016/j.jemermed.2018.12.025
251. Hoot NR, Aronsky D. Systematic Review of Emergency Department Crowding: Causes, Effects, and Solutions. *Ann Emerg Med*. 2008;52(2). doi:10.1016/j.annemergmed.2008.03.014
252. Asplin BR, Magid DJ, Rhodes K V., Solberg LI, Lurie N, Camargo CA. A conceptual model of emergency department crowding. *Ann Emerg Med*. 2003;42(2):173-180. doi:10.1067/mem.2003.302
253. Van Der Linden MC, Khursheed M, Hooda K, Pines JM, Van Der Linden N. Two emergency departments, 6000 km apart: Differences in patient flow and staff perceptions about crowding. *Int Emerg Nurs*. 2017;35:30-36. doi:10.1016/j.ienj.2017.06.002
254. Trzeciak S. *Emergency Department Overcrowding in the United States: An Emerging Threat to Patient Safety and Public Health*. Vol 20.; 2003. www.emjonline.com
255. Estey A, Ness K, Saunders LD, Alibhai A, Bear RA. Understanding the causes of overcrowding in emergency departments in the Capital Health Region in Alberta: A focus group study. *Canadian Journal of Emergency Medicine*. 2003;5(2):87-94. doi:10.1017/S1481803500008216
256. Aboagye-Sarfo P, Mai Q, Sanfilippo FM, Preen DB, Stewart LM, Fatovich DM. Growth in Western Australian emergency department demand during 2007-2013 is due to people with urgent and complex care needs. *EMA - Emergency Medicine Australasia*. 2015;27(3):202-209. doi:10.1111/1742-6723.12396

257. Moineddin R, Meaney C, Agha M, Zagorski B, Glazier RH. Modeling factors influencing the demand for emergency department services in ontario: A comparison of methods. *BMC Emerg Med*. 2011;11. doi:10.1186/1471-227X-11-13
258. Knapman M, Bonner A. Overcrowding in medium-volume emergency departments: Effects of aged patients in emergency departments on wait times for non-emergent triage-level patients. *Int J Nurs Pract*. 2010;16(3):310-317. doi:10.1111/j.1440-172X.2010.01846.x
259. Kawano T, Nishiyama K, Anan H, Tujimura Y. Direct relationship between aging and overcrowding in the ED, and a calculation formula for demand projection: A cross-sectional study. *Emergency Medicine Journal*. 2014;31(1):19-23. doi:10.1136/emermed-2012-202050
260. Cowling TE, Cecil E V., Soljak MA, et al. Access to Primary Care and Visits to Emergency Departments in England: A Cross-Sectional, Population-Based Study. *PLoS One*. 2013;8(6). doi:10.1371/journal.pone.0066699
261. Estey A, Ness K, Saunders LD, Alibhai A, Bear RA. Understanding the causes of overcrowding in emergency departments in the Capital Health Region in Alberta: A focus group study. *Canadian Journal of Emergency Medicine*. 2003;5(2):87-94. doi:10.1017/S1481803500008216
262. Derlet RW, Richards JR. Emergency department overcrowding in Florida, New York, and Texas. *South Med J*. 2002;95(8):846+. https://link.gale.com/apps/doc/A90569919/AONE?u=googlescholar&sid=googleScholar&xid=6a986a60
263. Kelen G, Scheulen JJ, Hill PM. Effect of an emergency department (ED) managed acute care unit on ED overcrowding and emergency medical services diversion. *Acad Emerg Med*. 2001;8 11:1095-1100.
264. Kenneth Bond Sandra Blitz Marc Afilalo Sam G. Campbell Michael Bullard Grant Innes Brian Holroyd Gil Curry Michael Schull and Brian H. Rowe MBO. Frequency, Determinants and Impact of Overcrowding in Emergency Departments in Canada: A National Survey. *Healthcare Quarterly*. 2007;10(4):32-40. https://www.longwoods.com/product/19312
265. Dunn R. Reduced Access Block Causes Shorter Emergency Department Waiting Times: An Historical Control Observational Study. *Emerg Med (Fremantle)*. 2003;15:232-238. doi:10.1046/j.1442-2026.2003.00441.x
266. Fatovich DM, Nagree Y, Sprivulis P. Access block causes emergency department overcrowding and ambulance diversion in Perth, Western Australia. *Emergency Medicine Journal*. 2005;22(5):351-354. doi:10.1136/emj.2004.018002
267. Forster AJ, Stiell I, Wells G, Lee AJ, Van Walraven C. The effect of hospital occupancy on emergency department length of stay and patient disposition. *Academic Emergency Medicine*. 2003;10(2):127-133. doi:10.1197/aemj.10.2.127
268. Lucas R, Farley HL, Twanmoh JR, et al. Emergency department patient flow: the influence of hospital census variables on emergency department length of stay. *Acad Emerg Med*. 2009;16 7:597-602.
269. Chalfin DB, Trzeciak S, Likourezos A, Baumann BM, Dellinger RP. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Crit Care Med*. 2007;35(6):1477-1483. doi:10.1097/01.CCM.0000266585.74905.5A

270. Sprivulis Peter C, Da Silva Julie-Ann, Jacobs Ian G, Frazer Amanda RL, Jelinek George A. The association between hospital overcrowding and mortality among patients admitted via western australian emergency departments. *Medical Journal of Australia*. 206AD;184(5):208-212. doi:10.5694/j.1326-5377.2006.tb00203.x
271. Stock LM, Bradley GE, Lewis J, Baker DW, Sipsey J, Stevens CD. Patients Who Leave Emergency Departments Without Being Seen by a Physician: Magnitude the Problem in Los Angeles County of. *Ann Emerg Med*. 1994;23(2):294-298. doi:https://doi.org/10.1016/S0196-0644(94)70043-5
272. Pines JM, Hollander JE, Localio AR, Metlay JP. The Association between Emergency Department Crowding and Hospital Performance on Antibiotic Timing for Pneumonia and Percutaneous Intervention for Myocardial Infarction. *Academic Emergency Medicine*. 2006;13(8):873-878. doi:10.1197/j.aem.2006.03.568
273. Adams J, Orav EJ, Rucker DW, Brennan TA, Burstin HR. Determinants of Patient Satisfaction and Willingness to Return With Emergency Care. *Ann Emerg Med*. 2000;35(5):426-434. doi:10.1067/mem.2000.104195
274. Krochmal P, Riley TA. Increased Health Care Costs Associated With ED Overcrowding. *Am J Emerg Med*. 1994;12(3):265-266. doi:10.1016/0735-6757(94)90135-X
275. Adriaenssens J, De Gucht V, Maes S. Determinants and prevalence of burnout in emergency nurses: A systematic review of 25 years of research. *Int J Nurs Stud*. 2015;52(2):649-661. doi:10.1016/j.ijnurstu.2014.11.004
276. Kulstad EB, Sikka R, Sweis RT, Kelley KM, Rzechula KH. ED overcrowding is associated with an increased frequency of medication errors. *American Journal of Emergency Medicine*. 2010;28(3):304-309. doi:10.1016/j.ajem.2008.12.014
277. Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. Emergency department crowding: A systematic review of causes, consequences and solutions. *PLoS One*. 2018;13(8). doi:10.1371/journal.pone.0203316
278. Kirubarajan A, Taher A, Khan S, Masood S. Artificial intelligence in emergency medicine: A scoping review. *J Am Coll Emerg Physicians Open*. 2020;1(6):1691-1702. doi:10.1002/emp2.12277
279. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. *NPJ Digit Med*. 2019;2(1):1-9. doi:10.1038/s41746-019-0111-3
280. Lewis TL, Wyatt JC. MHealth and mobile medical apps: A framework to assess risk and promote safer use. *J Med Internet Res*. 2014;16(9):1-7. doi:10.2196/jmir.3133
281. Kamel Boulos MN, Brewer AC, Karimkhani C, Buller DB, Dellavalle RP. Mobile medical and health apps: state of the art, concerns, regulatory control and certification. *Online J Public Health Inform*. 2014;5(3). doi:10.5210/ojphi.v5i3.4814
282. Hill MG, Sim M, Mills B. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. *Medical Journal of Australia*. 2020;212(11):514-519. doi:10.5694/mja2.50600
283. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: Audit study. *BMJ (Online)*. 2015;351. doi:10.1136/bmj.h3480
284. Chambers D, Cantrell AJ, Johnson M, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: Systematic review. *BMJ Open*. 2019;9(8). doi:10.1136/bmjopen-2018-027743

285. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage Accuracy of Symptom Checker Apps: 5-Year Follow-up Evaluation. *J Med Internet Res*. 2022;24(5). doi:10.2196/31810
286. Baker A, Perov Y, Middleton K, et al. A Comparison of Artificial Intelligence and Human Doctors for the Purpose of Triage and Diagnosis. *Front Artif Intell*. 2020;3. doi:10.3389/frai.2020.543405
287. Middleton K, Butt M, Hammerla N, Hamblin S, Mehta K, Parsa A. Sorting out symptoms: design and evaluation of the "babylon check" automated triage system. *ArXiv*. Published online 2016.
288. Bellika JG, Marco L, Wynn R. A Communicable Disease Query Engine. *Digital Healthcare Empowering Europeans*. Published online 2015:1012-1012. doi:10.3233/978-1-61499-512-8-1012
289. Arnold RJ, Layton A. Cost Analysis and Clinical Outcomes of Ambulatory Care Monitoring in Medicare Patients: Describing the Diagnostic Odyssey. *J Health Econ Outcomes Res*. 2013;2(2):161-169. doi:10.36469/9897
290. Miller S, Gilbert S, Virani V, Wicks P. Patients utilization and perception of an artificial intelligence based symptom assessment and advice technology in a British primary care waiting room: Exploratory pilot study. *JMIR Hum Factors*. 2020;7(3). doi:10.2196/19713
291. Armstrong S. The apps attempting to transfer NHS 111 online. *BMJ (Online)*. 2018;360. doi:10.1136/bmj.k156
292. Babylon Health. Accessed September 5, 2022. https://www.babylonhealth.com/
293. Ayanouz S, Abdelhakim BA, Benhmed M. A Smart Chatbot Architecture based NLP and Machine Learning for Health Care Assistance. In: *ACM International Conference Proceeding Series*. Association for Computing Machinery; 2020. doi:10.1145/3386723.3387897
294. Joseph Enguehard. Neural Temporal Point Processes (Neural TPPs). Published online 2020. Accessed September 21, 2022. https://github.com/babylonhealth/neuralTPPs
295. Enguehard J, Health B, Busbridge D, et al. *Neural Temporal Point Processes Modelling Electronic Health Records*. Vol 136.; 2020.
296. Pittet V, Burnand B, Yersin B, Carron PN. Trends of pre-hospital emergency medical services activity over 10 years: A population-based registry analysis. *BMC Health Serv Res*. 2014;14(1). doi:10.1186/1472-6963-14-380
297. Cabral ELDS, Castro WRS, Florentino DR de M, et al. Response time in the emergency services. Systematic review. *Acta Cir Bras*. 2018;33(12):1110-1121. doi:10.1590/s0102-865020180120000009
298. Lowthian JA, Cameron PA, Stoelwinder JU, et al. Increasing utilisation of emergency ambulances. *Australian Health Review*. 2011;35(1):63-69. doi:10.1071/AH09866
299. Byrsell F, Claesson A, Ringh M, et al. Machine learning can support dispatchers to better and faster recognize out-of-hospital cardiac arrest during emergency calls: A retrospective study. *Resuscitation*. 2021;162:218-226. doi:10.1016/j.resuscitation.2021.02.041
300. Borgholt L, Havtorn JD, Agić Ž, Søgaard A, Maaløe L, Igel C. Do End-to-End Speech Recognition Models Care About Context? *ArXiv*. Published online February 17, 2021. doi:10.21437/Interspeech.2020-1750
301. Miller M, Bootland D, Jorm L, Gallego B. Improving ambulance dispatch triage to trauma: A scoping review using the framework of development and evaluation of

clinical prediction rules. *Injury*. 2022;53(6):1746-1755. doi:10.1016/j.injury.2022.03.020
302. Naseem U, Dunn AG, Khushi M, Kim J. Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT. *BMC Bioinformatics*. 2022;23(1). doi:10.1186/s12859-022-04688-w
303. af Ugglas B, Skyttberg N, Wladis A, Djärv T, Holzmann MJ. Emergency department crowding and hospital transformation during COVID-19, a retrospective, descriptive study of a university hospital in Stockholm, Sweden. *Scand J Trauma Resusc Emerg Med*. 2020;28(1). doi:10.1186/s13049-020-00799-6
304. Saberian P, Conovaloff JL, Vahidi E, Hasani-Sharamin P, Kolivand PH. How the COVID-19 epidemic affected prehospital emergency medical services in Tehran, Iran. *Western Journal of Emergency Medicine*. 2020;21(6). doi:10.5811/WESTJEM.2020.8.48679
305. Paudel P. 911 Overflow. Published 2019. Accessed September 5, 2022. https://devpost.com/software/911-overflow
306. Pfeifer R, Halvachizadeh S, Schick S, et al. Are Pre-hospital Trauma Deaths Preventable? A Systematic Literature Review. *World J Surg*. 2019;43(10):2438-2446. doi:10.1007/s00268-019-05056-1
307. Tollinton L, Metcalf AM, Velupillai S. Enhancing predictions of patient conveyance using emergency call handler free text notes for unconscious and fainting incidents reported to the London Ambulance Service. *Int J Med Inform*. 2020;141. doi:10.1016/j.ijmedinf.2020.104179
308. Ferri P, Sáez C, Félix-De Castro A, et al. Deep ensemble multitask classification of emergency medical call incidents combining multimodal data improves emergency medical dispatch. *Artif Intell Med*. 2021;117. doi:10.1016/j.artmed.2021.102088
309. Fix J, Ising AI, Proescholdbell SK, et al. Linking Emergency Medical Services and Emergency Department Data to Improve Overdose Surveillance in North Carolina. *Public Health Methodology Public Health Reports*. 2021;136:54-61. doi:10.1177/00333549211012400
310. Martin TJ, Ranney ML, Dorroh J, Asselin N, Sarkar IN. Health information exchange in emergency medical services. *Appl Clin Inform*. 2018;9(04):884-891.
311. Redfield C, Tlimat A, Halpern Y, et al. Derivation and validation of a machine learning record linkage algorithm between emergency medical services and the emergency department. *Journal of the American Medical Informatics Association*. 2020;27(1):147-153. doi:10.1093/jamia/ocz176
312. Kirkland SW, Soleimani A, Rowe BH, Newton AS. A systematic review examining the impact of redirecting low-acuity patients seeking emergency department care: Is the juice worth the squeeze? *Emergency Medicine Journal*. 2019;36(2):97-106. doi:10.1136/emermed-2017-207045
313. Worster A, Fernandes CM, Eva K, Upadhye S. Predictive validity comparison of two five-level triage acuity scales. *European Journal of Emergency Medicine*. 2007;14:188-192.
314. Farrohknia N, Castrén M, Ehrenberg A, et al. Emergency Department Triage Scales and Their Components: A Systematic Review of the Scientific Evidence. *Scand J Trauma Resusc Emerg Med*. 2011;19. doi:10.1186/1757-7241-19-42
315. Christ M, Grossmann F, Winter D, Bingisser R, Platz E. Modern triage in the emergency department. *Dtsch Arztebl Int*. 2010;107(50):892.

316. Hinson JS, Martinez DA, Cabral S, et al. Triage Performance in Emergency Medicine: A Systematic Review. *Ann Emerg Med*. 2019;74(1):140-152. doi:10.1016/j.annemergmed.2018.09.022
317. Fernandes M, Vieira SM, Leite F, Palos C, Finkelstein S, Sousa JMC. Clinical Decision Support Systems for Triage in the Emergency Department using Intelligent Systems: a Review. *Artif Intell Med*. 2020;102. doi:10.1016/j.artmed.2019.101762
318. Levin S, Toerper M, Hamrock E, et al. Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index. *Ann Emerg Med*. 2018;71(5):565-574.e2. doi:10.1016/j.annemergmed.2017.08.005
319. Ivanov O, Wolf L, Brecher D, et al. Improving ED Emergency Severity Index Acuity Assignment Using Machine Learning and Clinical Natural Language Processing. *J Emerg Nurs*. 2021;47(2):265-278.e7. doi:10.1016/j.jen.2020.11.001
320. Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Ann Intern Med*. 2016;165(11):753-760. doi:10.7326/M16-0961
321. Gardner RL, Cooper E, Haskell J, et al. Physician stress and burnout: the impact of health information technology. *Journal of the American Medical Informatics Association*. 2019;26(2):106-114. doi:10.1093/jamia/ocy145
322. Carayon P, Wetterneck TB, Alyousef B, et al. Impact of electronic health record technology on the work and workflow of physicians in the intensive care unit. *Int J Med Inform*. 2015;84(8):578-594. doi:10.1016/j.ijmedinf.2015.04.002
323. Chan KS, Fowles JB, Weiner JP. Electronic health records and the reliability and validity of quality measures: a review of the literature. *Medical Care Research and Review*. 2010;67(5):503-527.
324. Kossman SP, Scheidenhelm SL. Nurses' Perceptions of the Impact of Electronic Health Records on Work and Patient Outcomes. *Comput Inform Nurs*. 2008;26(2):69-77. doi:10.1097/01.NCN.0000304775.40531.67
325. Greenbaum NR, Jernite Y, Halpern Y, et al. Contextual Autocomplete: A Novel User Interface Using Machine Learning to Improve Ontology Usage and Structured Data Capture for Presenting Problems in the Emergency Department. *MIT Clinical*. Published online 2017. doi:10.1101/127092
326. Murray L, Gopinath Di, Agrawal M, Horng S, Sontag D, Karger DR. MedKnowts: Unified Documentation and Information Retrieval for Electronic Health Records. In: *UIST 2021 - Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, Inc; 2021:1169-1183. doi:10.1145/3472749.3474814
327. Li J, Chen X, Gaur Y, et al. Recent Advances in End-to-End Automatic Speech Recognition. *APSIPA Trans Signal Inf Process*. 2022;11:8. doi:10.1561/116.00000050_supp
328. Academy of Family Physicians A. *Using an AI Assistant To Reduce Documentation Burden in Family Medicine Evaluating the Suki Assistant*.; 2021. Accessed October 11, 2022. https://www.aafp.org/dam/AAFP/documents/practice_management/innovation_lab/report-suki-assistant-documentation-burden.pdf

329. Dragon Medical One. Accessed September 21, 2022. https://www.nuance.com/healthcare/provider-solutions/speech-recognition/dragon-medical-one.html
330. Elayan H, Aloqaily M, Guizani M. Digital Twin for Intelligent Context-Aware IoT Healthcare Systems. *IEEE Internet Things J*. 2021;8(23):16749-16757. doi:10.1109/JIOT.2021.3051158
331. Chaou CH, Chen HH, Chang SH, et al. Predicting length of stay among patients discharged from the emergency department-using an accelerated failure time model. *PLoS One*. 2017;12(1). doi:10.1371/journal.pone.0165756
332. Hojat M, Louis DZ, Markham FW, Wender R, Rabinowitz C, Gonnella JS. Physicians' empathy and clinical outcomes for diabetic patients. *Academic Medicine*. 2011;86(3):359-364. doi:10.1097/ACM.0b013e3182086fe1
333. Kelley JM, Kraft-Todd G, Schapira L, Kossowsky J, Riess H. The influence of the patient-clinician relationship on healthcare outcomes: A systematic review and meta-analysis of randomized controlled trials. *PLoS One*. 2014;9(4). doi:10.1371/journal.pone.0094207
334. Richter JP, Muhlestein DB. Patient experience and hospital profitability: Is there a link? *Health Care Manage Rev*. 2017;42(3):247-257. doi:10.1097/HMR.0000000000000105
335. Pitrou I, Lecourt AC, Bailly L, Brousse B, Dauchet L, Ladner J. Waiting time and assessment of patient satisfaction in a large reference emergency department: A prospective cohort study, France. *European Journal of Emergency Medicine*. 2009;16(4):177-182. doi:10.1097/MEJ.0b013e32831016a6
336. Sonis JD, Aaronson EL, Lee RY, Philpotts LL, White BA. Emergency Department Patient Experience. *J Patient Exp*. 2018;5(2):101-106. doi:10.1177/2374373517731359
337. Crilly J, Chaboyer W, Creedy D. Violence towards emergency department nurses by patients. *Accid Emerg Nurs*. 2004;12(2):67-73. doi:10.1016/j.aaen.2003.11.003
338. Pich J, Hazelton M, Sundin D, Kable A. Patient-related violence at triage: A qualitative descriptive study. *Int Emerg Nurs*. 2011;19(1):12-19. doi:10.1016/j.ienj.2009.11.007
339. Krishel S, Larry Baraff MJ. Effect of Emergency Department Information on Patient Satisfaction. *Ann Emerg Med*. 1993;22(3):568-572. doi:10.1016/S0196-0644(05)81943-2
340. Hassan R, Twynam NW, Nah FF,, Siau K. Patient engagement in the medical facility waiting room using Gamified healthcare information delivery. *International Conference on HCI in Businness, Government, and Organizations*.:412-423. doi:10.1007/978-3-319-39399-5_39
341. Göransson KE, von Rosen A. Patient experience of the triage encounter in a Swedish emergency department. *Int Emerg Nurs*. 2010;18(1):36-40. doi:10.1016/j.ienj.2009.10.001
342. Broida RI, Desai SA, Easter BD, et al. *Emergency Department Crowding: Emergency Medicine Practice Committee High Impact Solutions Subcommittee Members*.; 2016. https://www.acep.org/globalassets/sites/acep/media/crowding/empc_crowding-ip_092016.pdf
343. Xie C, Zhang J, Morrison AM, Coca-Stefaniak JA. The effects of risk message frames on post-pandemic travel intentions: The moderation of empathy and perceived waiting time. *Current Issues in Tourism*. 2021;24(23):3387-3406. doi:10.1080/13683500.2021.1881052

344. Kilaru AS, Meisel ZF, Paciotti B, et al. What do patients say about emergency departments in online reviews? A qualitative study. *BMJ Qual Saf*. 2016;25(1):14-24. doi:10.1136/bmjqs
345. Al-Nerabieah Z, Alhalabi MN, Owayda A, Alsabek L, Bshara N, Kouchaji C. Effectiveness of using virtual reality eyeglasses in the waiting room on preoperative anxiety: A Randomized Controlled Trial. *Perioper Care Oper Room Manag*. 2020;21. doi:10.1016/j.pcorm.2020.100129
346. Dandu K V., Carniol ET, Sanghvi S, Baredes S, Eloy JA. A 10-Year Analysis of Head and Neck Injuries Involving Nonpowder Firearms. *Otolaryngology - Head and Neck Surgery (United States)*. 2017;156(5):853-856. doi:10.1177/0194599817695546
347. Josseran L, Fouillet A, Caillère N, et al. Assessment of a syndromic surveillance system based on morbidity data: Results from the Oscour® network during a heat wave. *PLoS One*. 2010;5(8). doi:10.1371/journal.pone.0011984
348. Gil-Jardiné C, Chenais G, Pradeau C, et al. Surveillance of COVID-19 using a keyword search for symptoms in reports from emergency medical communication centers in Gironde, France: a 15 year retrospective cross-sectional study. *Intern Emerg Med*. 2022;17(2):603-608. doi:10.1007/s11739-021-02818-5
349. Sahu KS, Majowicz SE, Dubin JA, Morita PP. NextGen Public Health Surveillance and the Internet of Things (IoT). *Front Public Health*. 2021;9. doi:10.3389/fpubh.2021.756675
350. National Institute of Standards and Technology. *AI Risk Management Framework: Second Draft Notes for Reviewers: Call for Comments and Contributions*.; 2022. Accessed September 22, 2022. https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf
351. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health*.; 2021. http://apps.who.int/bookorders.
352. Meskó B, Spiegel B. A Revised Hippocratic Oath for The Era of Digital Health (Preprint). *J Med Internet Res*. Published online September 7, 2022. doi:10.2196/39177
353. International Organization for Standardization. *ISO/IEC TS 5723:2022 (3.2.17)*.; 2022. Accessed September 23, 2022. https://www.iso.org/obp/ui/fr/#iso:std:iso-iec:ts:5723:ed-1:v1:en
354. Zhuang S, Hadfield-Menell D. Consequences of Misaligned AI. In: *34th Conference on Neural Information Processing Systems (NeurIPS)*. ; 2020.
355. D'Amour A, Heller K, Moldovan D, et al. Underspecification Presents Challenges for Credibility in Modern Machine Learning. Published online November 6, 2020. http://arxiv.org/abs/2011.03395
356. Passi S, Barocas S. Problem formulation and fairness. In: *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc; 2019:39-48. doi:10.1145/3287560.3287567
357. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete Problems in AI Safety. Published online June 21, 2016. http://arxiv.org/abs/1606.06565
358. Barocas S, Hardt M, Narayanan A. Fairness and Machine Learning Limitations and Opportunities. *ACM Comput Surv*. 2002;54(6). doi:10.1145/3457607
359. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput Surv*. 2021;54(6). doi:10.1145/3457607
360. Milner KA, Vaccarino V, Arnold AL, Funk M, Goldberg RJ. Gender and age differences in chief complaints of acute myocardial infarction (Worcester Heart Attack Study).

*American Journal of Cardiology*. 2004;93(5):606-608. doi:10.1016/j.amjcard.2003.11.028
361. Bozkurt B, Khalaf S. Heart Failure in Women. *Methodist Debakey Cardiovasc J*. 2017;13(4):216-223. doi:10.14797/mdcj-13-4-216
362. Suresh H, Guttag J V. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. Published online January 28, 2019. doi:10.1145/3465416.3483305
363. Blyth CR. On Simpson's Paradox and the Sure-Thing Principle. *Source: Journal of the American Statistical Association*. 1972;67(338):364-366. doi:10.2307/2284382
364. Forbes LA, Canner JK, Milio L, Halscott T, Vaught AJ. Association of Patient Sex and Pregnancy Status With Naloxone Administration During Emergency Department Visits. *Obstetrics and gynecology*. 2021;137(5):855-863. doi:10.1097/AOG.0000000000004357
365. Gehlke CE, Biehl K. Certain Effects of Grouping upon the Size of the Correlation Coefficient in Census Tract Material. *J Am Stat Assoc*. 1934;29(185A):169-170. doi:10.1080/01621459.1934.10506247
366. Kok MR, Tuson M, Yap M, et al. Impact of the modifiable areal unit problem in assessing determinants of emergency department demand. *EMA - Emergency Medicine Australasia*. 2021;33(5):794-802. doi:10.1111/1742-6723.13727
367. Reimer AP, Milinovich A, Madigan EA. Data quality assessment framework to assess electronic medical record data for use in research. *Int J Med Inform*. 2016;90:40-47. doi:10.1016/j.ijmedinf.2016.03.006
368. Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of In-hospital Mortality in Emergency Department Patients with Sepsis: A Local Big Data-Driven, Machine Learning Approach. *Academic Emergency Medicine*. 2016;23(3):269-278. doi:10.1111/acem.12876
369. Barberà-Mariné MG, Cannavacciuolo L, Ippolito A, Ponsiglione C, Zollo G. The weight of organizational factors on heuristics: Evidence from triage decision-making processes. *Management Decision*. 2019;57(11):2890-2910. doi:10.1108/MD-06-2017-0574
370. Mohan D, Fischhoff B, Angus DC, et al. Serious games may improve physician heuristics in trauma triage. *Proc Natl Acad Sci U S A*. 2018;115(37):9204-9209. doi:10.1073/pnas.1805450115
371. Hovy D, Prabhumoye S. Five sources of bias in natural language processing. *Lang Linguist Compass*. 2021;15(8). doi:10.1111/lnc3.12432
372. Paullada A, Raji ID, Bender EM, Denton E, Hanna A. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*. 2021;2(11). doi:10.1016/j.patter.2021.100336
373. Schrader CD, Lewis LM. Racial disparity in Emergency Department triage. *Journal of Emergency Medicine*. 2013;44(2):511-518. doi:10.1016/j.jemermed.2012.05.010
374. Arslanian-Engoren C. Gender and age bias in triage decisions. *J Emerg Nurs*. 2000;26(2):117-124. doi:10.1067/men.2000.105643
375. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc; 2021:610-623. doi:10.1145/3442188.3445922
376. Wagner C, Garcia D, Gesis MJ, Strohmaier M. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In: *Ninth International AAAI Conference on Web*

*and Social Media* . ; 2021:454-463. https://ojs.aaai.org/index.php/ICWSM/article/view/14628

377. Liang PP, Li IM, Zheng E, Lim YC, Salakhutdinov R, Morency LP. Towards Debiasing Sentence Representations. Published online July 16, 2020. http://arxiv.org/abs/2007.08100
378. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI Extension. *The BMJ*. 2020;370. doi:10.1136/bmj.m3164
379. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351-1363. doi:10.1038/s41591-020-1037-7
380. Xavie r S, Christophe G, Julien P, et al. *Intellectual Property and Digital Trade in the Age of Artificial Intelligence and Big Data Global Perspectives and Challenges for the Intellectual Property System*.; 2018. Accessed September 20, 2022. https://www.i3pm.org/files/misc/CEIPI-ICTSD_Issue_5.pdf
381. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In: *Artificial Intelligence in Healthcare*. Elsevier; 2020:295-336. doi:10.1016/B978-0-12-818438-7.00012-5
382. Maliha G, Gerke S, Cohen IG, Parikh RB. Artificial Intelligence and Liability in Medicine: Balancing Safety and Innovation. *Milbank Q*. 2021;00(0):1-19. doi:10.1111/1468-0009.12504
383. Price WN, Gerke S, Cohen IG. Potential Liability for Physicians Using Artificial Intelligence. *JAMA - Journal of the American Medical Association*. 2019;322(18):1765-1766. doi:10.1001/jama.2019.15064
384. Antoniadi AM, Du Y, Guendouz Y, et al. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review. *Applied Sciences (Switzerland)*. 2021;11(11). doi:10.3390/app11115088
385. Danilevsky M, Qian K, Aharonov R, Katsis Y, Kawas B, Sen P. A Survey of the State of Explainable AI for Natural Language Processing. Published online October 1, 2020. http://arxiv.org/abs/2010.00711
386. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*. 2020;58:82-115. doi:10.1016/j.inffus.2019.12.012
387. Castelvecchi D. Can we open the black box of AI? *Nature News*. 2016;538(7623):20. doi:10.1038/538020a
388. San Francisco Calif: Bancroft-Whitney; Deering's California Codes Annotated. . Published online 1957.
389. Moore G, Matlock A, Kiley J, Percy K. Emergency Physicians: Beware of the Consent Standard of Care. *Clin Pract Cases Emerg Med*. 2018;2(2):109-111. doi:10.5811/cpcem.2018.1.37822
390. European Comission. *Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive)*.; 2022. Accessed May 11, 2023. https://commission.europa.eu/system/files/2022-09/1_1_197605_prop_dir_ai_en.pdf

391. European Commission. *New Liability Rules on Products and AI to Protect Consumers and Foster Innovation.*; 2022. Accessed May 11, 2023. https://ec.europa.eu/commission/presscorner/detail/en/ip_22_5807
392. Forcier MB, Gallois H, Mullan S, Joly Y. Integrating artificial intelligence into health care through data access: Can the GDPR act as a beacon for policymakers? *J Law Biosci*. 2019;6(1):317-335. doi:10.1093/jlb/lsz013
393. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med*. 2018;15(11):e1002689.
394. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2019;10(2):1-19.
395. Gu R, Niu C, Wu F, et al. From server-based to client-based machine learning: a comprehensive survey. *ACM Computing Surveys (CSUR)*. 2021;54(1):1-36.
396. Antunes RS, da Costa C, Küderle A, Yari IA, Eskofier B. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2022;13(4):1-23.
397. NVIDIA blog. Accessed May 10, 2023. https://blogs.nvidia.com/blog/2019/10/13/what-is-federated-learning/
398. Grant K, McParland A, Mehta S, Ackery AD. Artificial Intelligence in Emergency Medicine: Surmountable Barriers With Revolutionary Potential. *Ann Emerg Med*. 2020;75(6):721-726. Doi:10.1016/j.annemergmed.2019.12.024
399. FEDORU. *RPU 02 – Format Des Éléments Collectés et Règles de Codage*.; 2016. Accessed May 9, 2021. https://fedoru.fr/fiche-publications/02-format-elements-collectes-regles-codage/

# VII. PUBLICATIONS

- Chenais, G, Gil-Jardiné C, Touchais H, Avalos Fernandez M, Contrand B, Tellier E, Combes X, Bourdois L, Revel P, Lagarde E, Deep Learning Transformer Models for Building a Comprehensive and Real-time Trauma Observatory: Development and Validation Study. *JMIR AI* 2023;2:e4084.
- Chenais G, Lagarde E, Gil-Jardiné C, Artificial Intelligence in Emergency Medicine: Viewpoint of Current Applications and Foreseeable Opportunities and Challenges. *J Med Internet Res* 2023;25:e40031
- Chenais, G., Touchais, H., Avalos, M., Bourdois, L., Revel, P., Gil-Jardiné, C., & Lagarde, E. (2021, June). Performance en classification de données textuelles des passages aux urgences des modèles BERT pour le français. In *PFIA 2021-Journée Santé et IA*.
- Chenais, G, Benchmarking Natural Language Processing tools for automatic classification of Trauma mechanism in French emergency free-text clinical notes. In *EU Safety 2022 Vienne*, Paper ID 41.

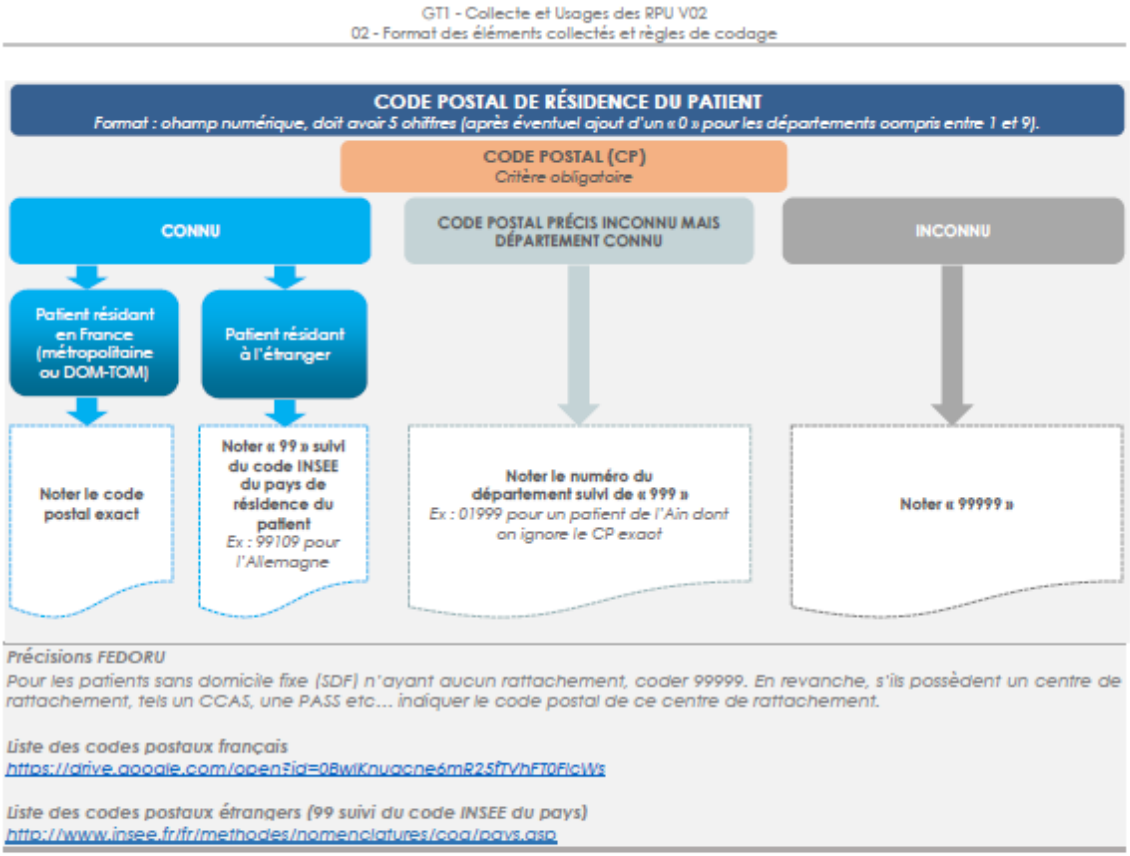
- Gil-Jardiné, C., Chenais, G., Pradeau, C., Tentillier, E., Revel, P., Combes, X., ... & Lagarde, E. (2022). Surveillance of COVID-19 using a keyword search for symptoms in reports from emergency medical communication centers in Gironde, France: a 15-year retrospective cross-sectional study. *Internal and Emergency Medicine*, *17*(2), 603-608.
- Gil-Jardiné, C., Chenais, G., Pradeau, C., Tentillier, E., Revel, P., Combes, X., ... & Lagarde, E. (2021). Trends in reasons for emergency calls during the COVID-19 crisis in the department of Gironde, France using artificial neural network for natural language classification. *Scandinavian journal of trauma, resuscitation and emergency medicine*, *29*, 1-9.
- Bourdois, L., Avalos, M., Chenais, G., Contrand, B., Gil-Jardiné, C., Guennec-Jacques, A., ... & Lagarde, E. (2021, June). Traitement automatique des résumés de passages aux urgences: focus sur la désidentification. In *PFIA 2021-Santé et IA*.
- Bourdois, L., Avalos, M., Chenais, G., Thiessard, F., Revel, P., Gil-Jardiné, C., & Lagarde, E. (2021). De-identification of emergency medical records in French: Survey and comparison of state-of-the-art automated systems. *Florida Artificial Intelligence Research Society*, *34*(1).

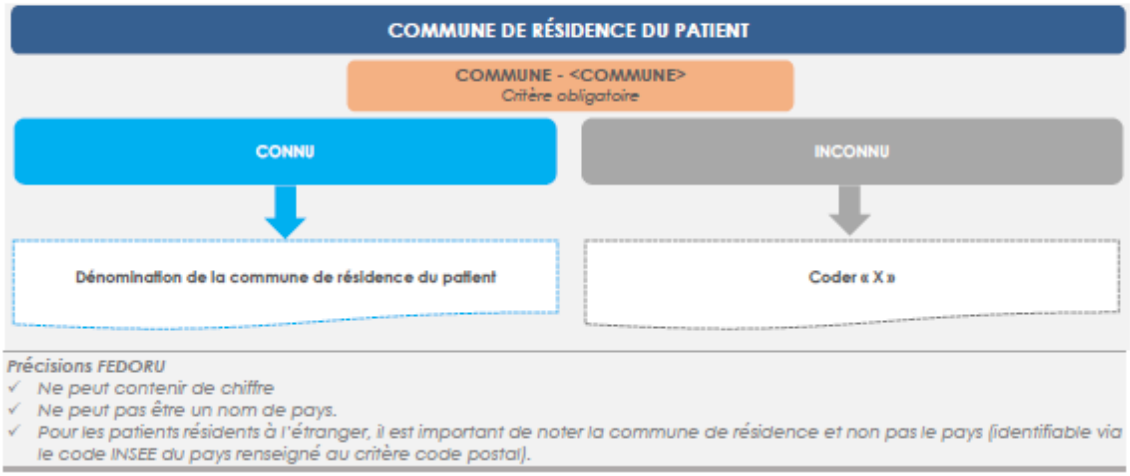
## Appendix A Résumé de Passage aux Urgences v2, variable definition and format

| TAG | FORMAT | DEFINITION |
|---|---|---|
| <ZC> | Integer | Zip Code of residency |
| < City> | Text | City name of residency |
| <BIRTH> | DD/MM/YYYY | Date of Birth (blank means uncertain) |
| <SEXE> | M / F / I | Sex (I for Undetermined) |
| <ADMISSION> | DD/MM/YYYY HH:MM | Date and Time for Admission |
| <ADMISSION MODE> | 6 / 7 / 8 | Admission Mode PMSI |
| <PROVENANCE> | 1/ 2 / 3 / 4 / 5 / 8 | Provenance Mode PMSI |
| <TRANSPORT> | PERRS / AMBU / VSAB / SMUR / HELI / FO | Type of Transport |
| < TRANSPORT CARE> | MED / PARAMED / AUCUN | Type of Care during Transportation |
| <MOTIVE> | THESAURUS SFMU | Motive for Emergency Visit SFMU |
| < SEVERITY> | 1 / 2 / 3 / 4 / 5 /P / D | CCMU Classification |
| <MD> | ICD-10 code | Main DIagnosis |
| <SD_LIST> <SD> </SD> | ICD-10 code | Secondary Diagnosis |
| <ACT_LIST> <ACT> </ACT> | CCAM code | Medical Act in Emergency Departement |
| <DISCHARGE> | DD//MM/YYYY HH:MM | Date and Time for Discharge |
| <DISCHARGE MODE> | 6 / 7 / 8 / 9 | Discharge Mode PMSI |
| <DESTINATION> | 1 / 2 / 3 / 4 / 5 / 6 / 7 | Destination PMSI |
| <ORIENTATION> | FUGUE / SCAM / PAS / REO / SC / SI / REA / UHCD / MED / CHIR / OBST / HDT / HO | Orientation Precision |

Back to section French emergency surveillance system

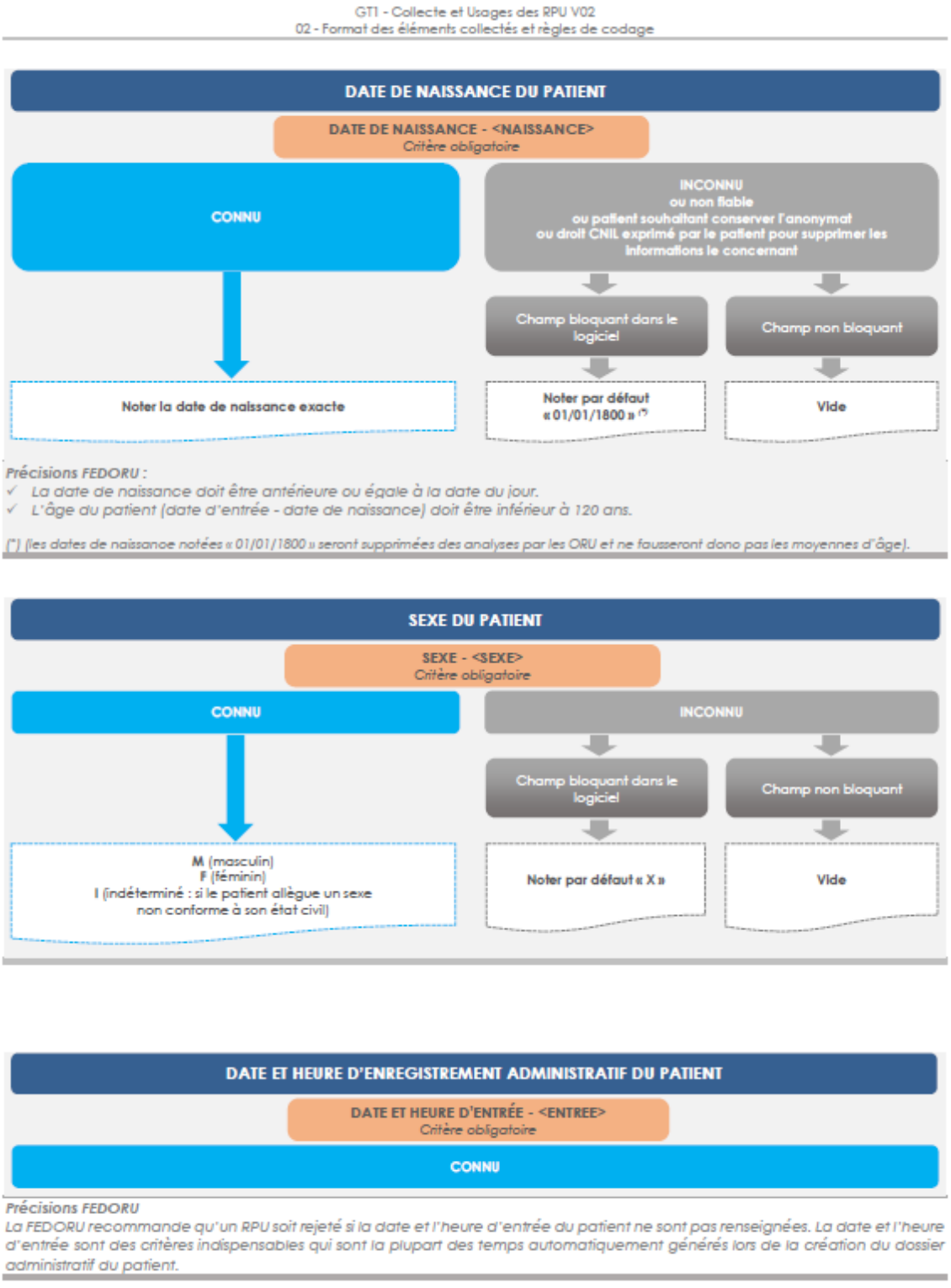
# Appendix B Résumé de Passage aux Urgences v2, format of collected data and coding rules[399]



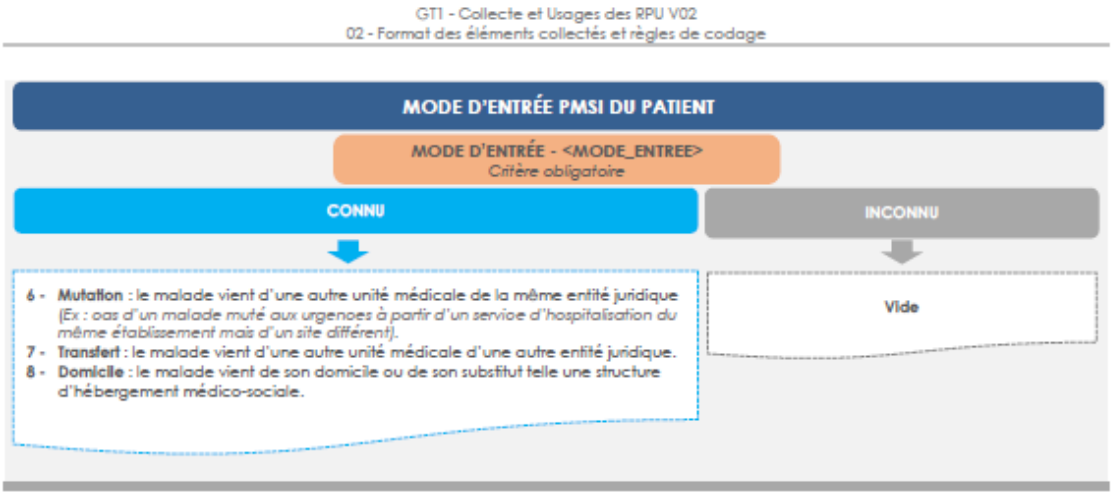

Back to section French emergency surveillance system

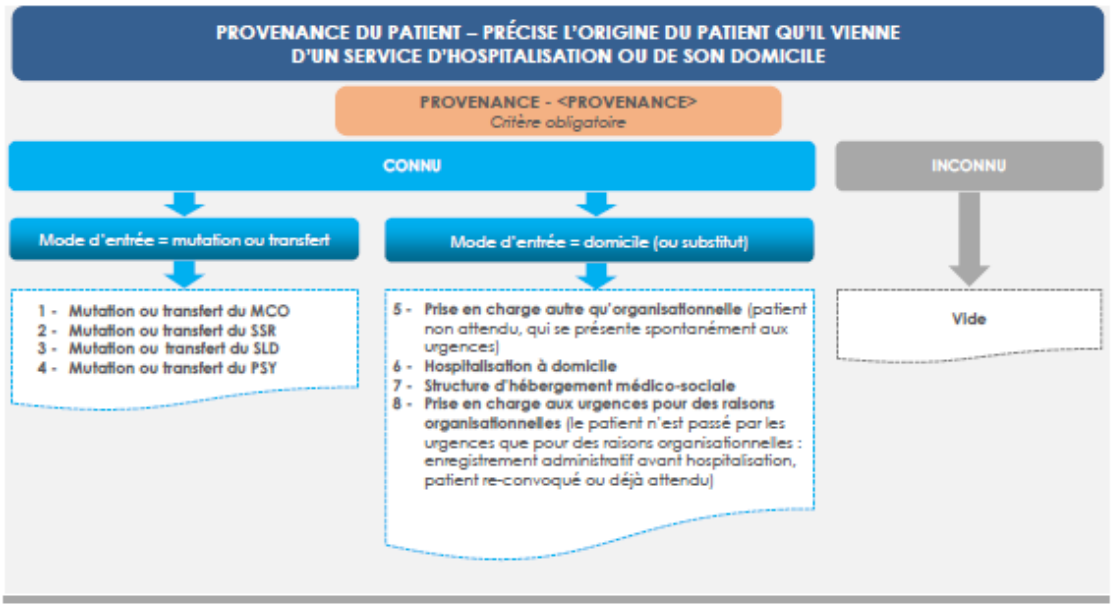
# Appendix C Résumé de Passage aux Urgences v2, format of collected data and coding rules[399]




Back to section French emergency surveillance system

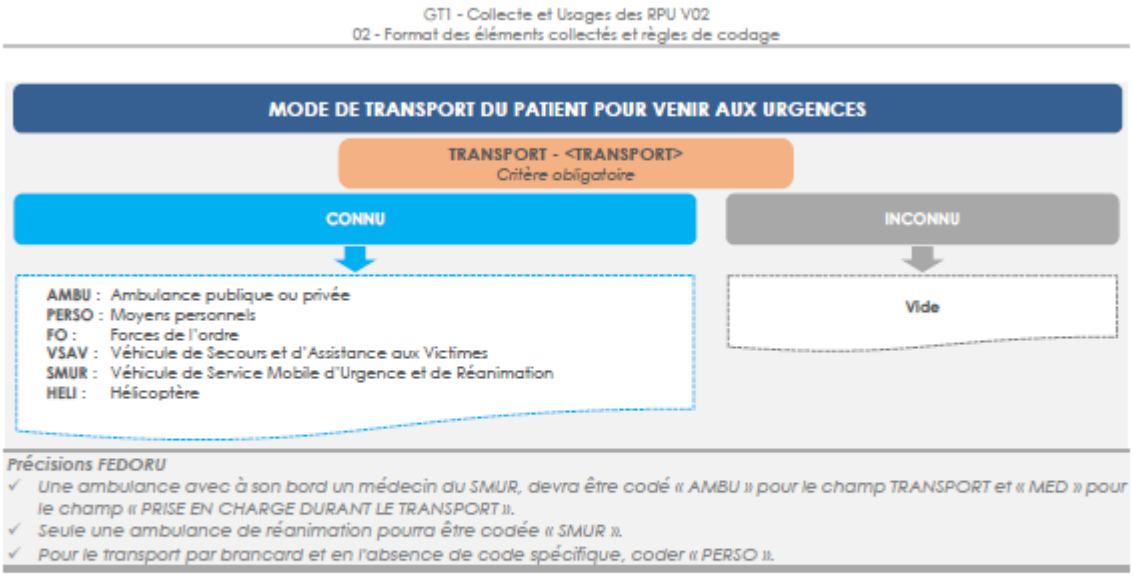
# Appendix D Résumé de Passage aux Urgences v2, format of collected data and coding rules[399]

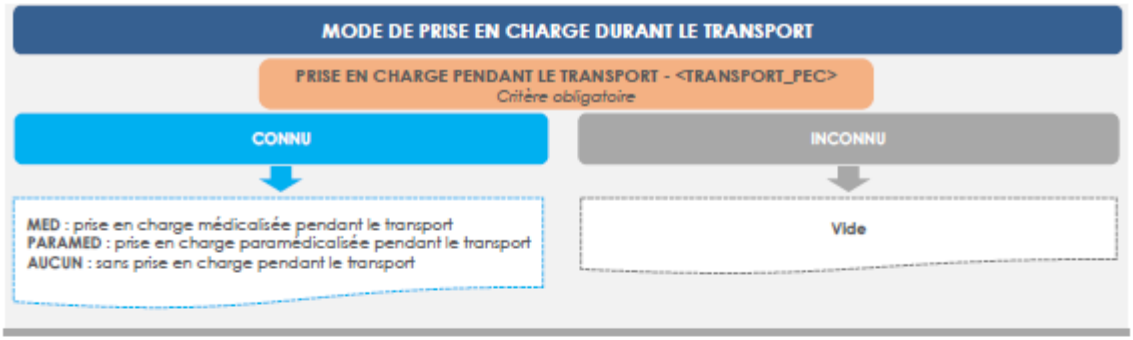


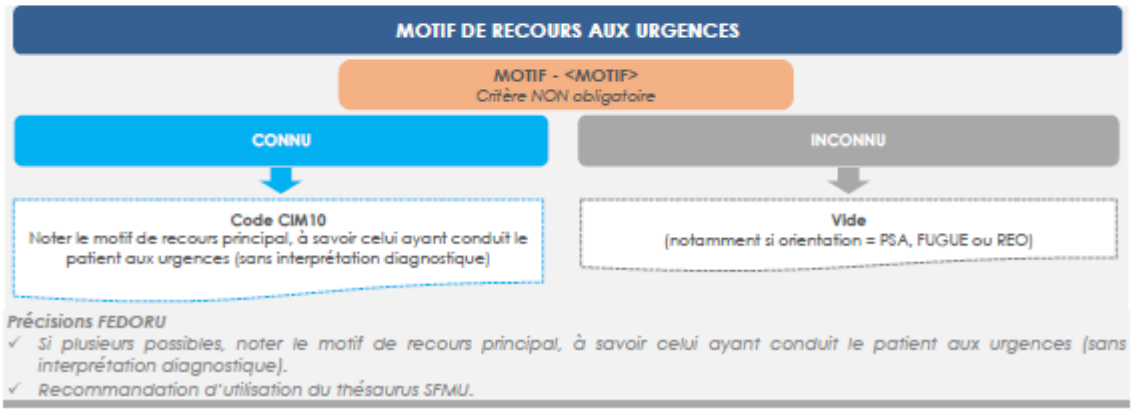




Back to section French emergency surveillance system

# Appendix E Résumé de Passage aux Urgences v2, format of collected data and coding rules[399]

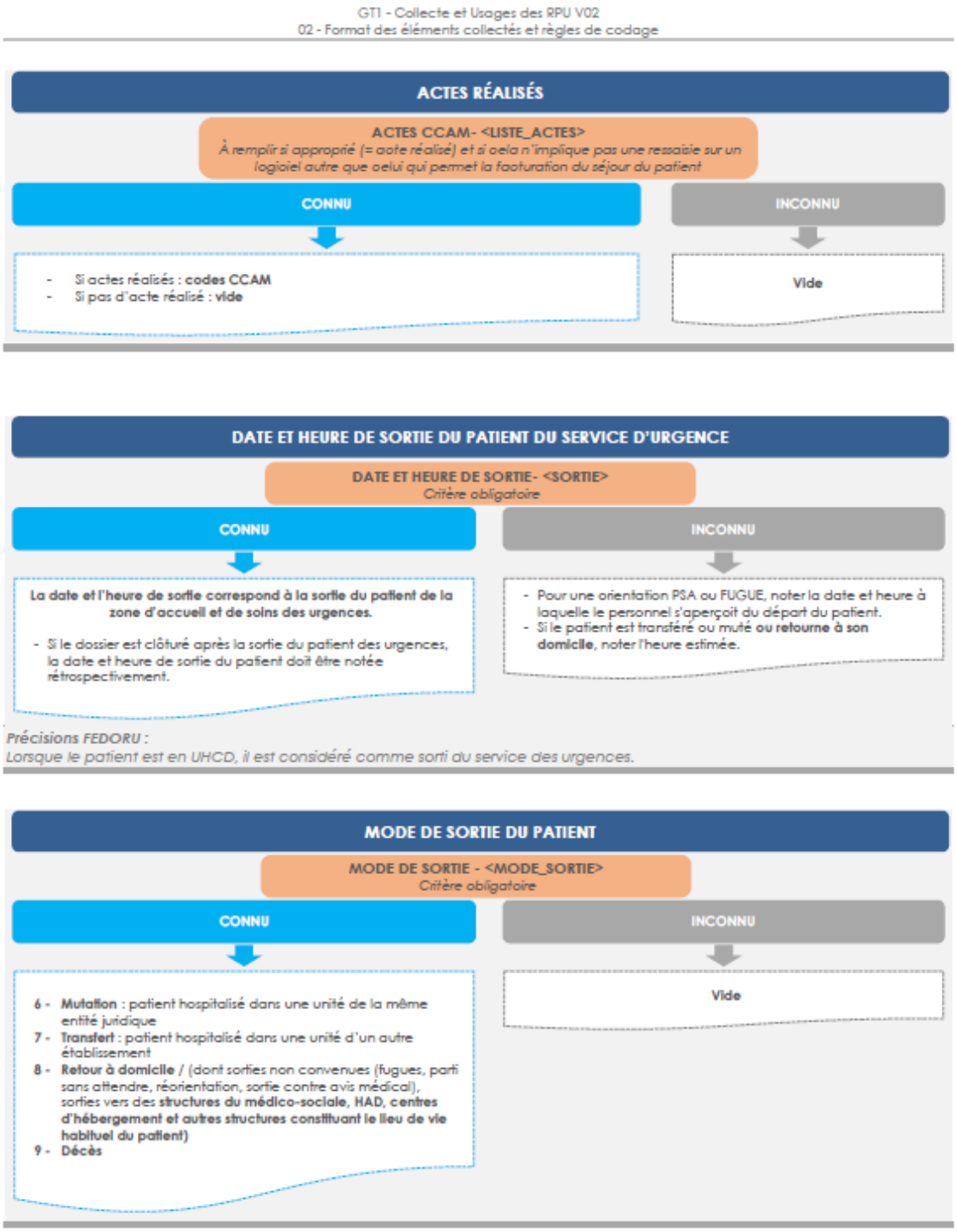


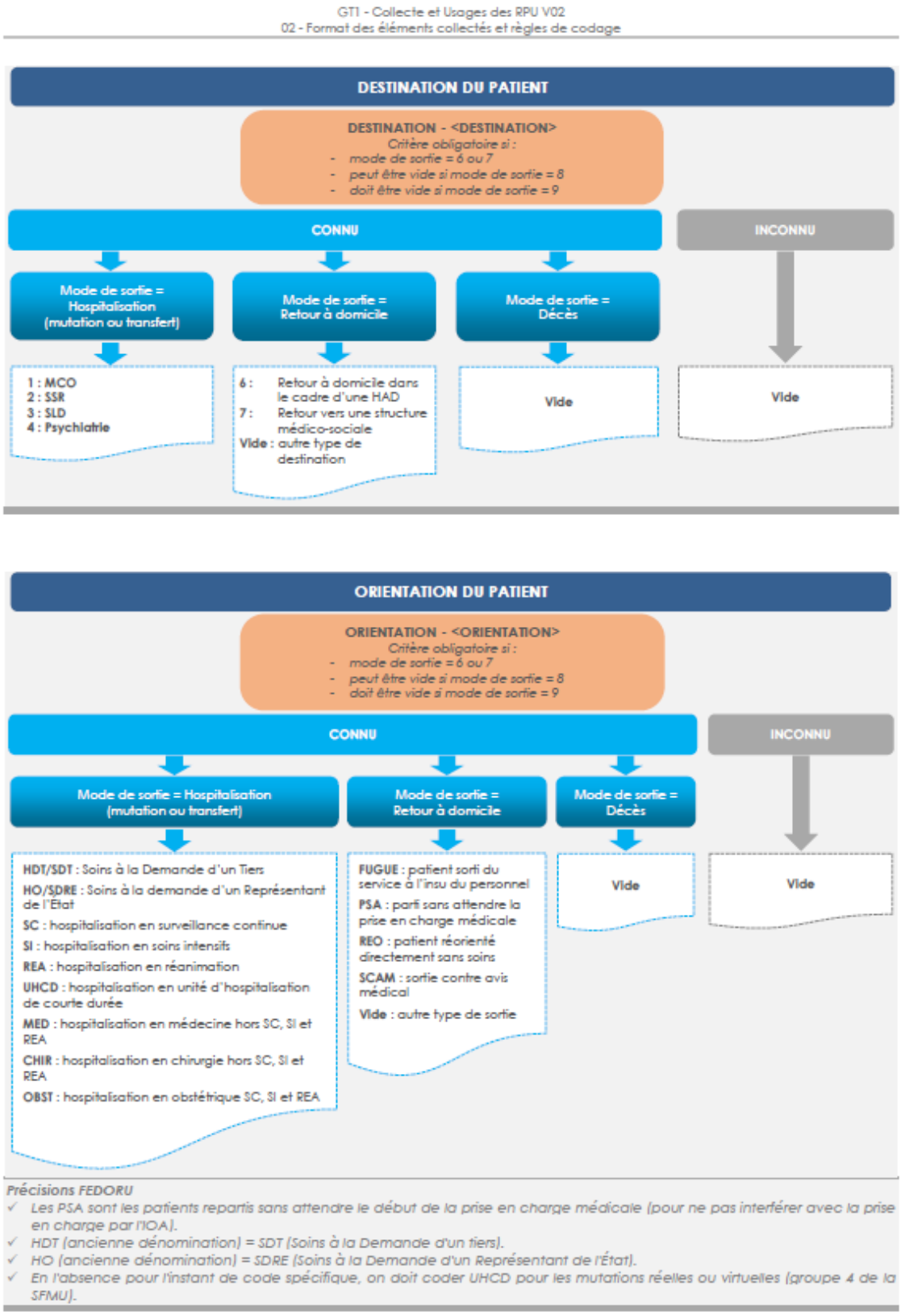

Back to section French emergency surveillance system

# Appendix F Résumé de Passage aux Urgences v2, format of collected data and coding rules[399]




Back to section French emergency surveillance system

# Appendix G Résumé de Passage aux Urgences v2, format of collected data and coding rules[399]




Back to section French emergency surveillance system

# Appendix H Ratios between total number of BAAC (law enforcement) injuries and Gustave Eiffel University estimate for MAIS1-2 and MAIS3+ on average over 2012-2016

## Méthode d'estimation du nombre de blessés[1]

| | |
|---|---|
| Équipe projet | ONISR |
| Partenaires | Emmanuelle Amoros (Univ. Eiffel) |
| Méthodologie | Définir des coefficients correcteurs simples entre le BAAC et l'estimation des blessés modélisée par l'Univ. Eiffel |
| Mots clés | Estimation, redressement, Registre du Rhône, nombre de blessés |

**Ratios entre le nombre total de blessés BAAC et l'estimation de l'Université Gustave Eiffel pour les MAIS1-2 et les MAIS3+ en moyenne sur 2012-2016**

| MAIS1-2 GN | Piéton | Vélo | EDPm | 2RM | VT | Autre |
|---|---|---|---|---|---|---|
| 0-13 ans | 2,82 | 17,56 | 17,56 | 7,48 | 3,35 | 2,29 |
| 14-19 ans | 2,47 | 15,60 | 15,60 | 5,75 | 5,40 | 2,84 |
| 20-29 ans | 2,19 | 10,45 | 10,45 | 3,76 | 4,44 | 2,31 |
| 30-39 ans | 1,88 | 10,68 | 10,68 | 2,82 | 3,89 | 1,96 |
| 40-49 ans | 1,99 | 11,26 | 11,26 | 2,89 | 3,84 | 1,91 |
| 50-59 ans | 1,98 | 8,94 | 8,94 | 2,54 | 3,39 | 1,78 |
| 60-69 ans | 1,67 | 7,22 | 7,22 | 2,32 | 2,89 | 1,43 |
| 70 ans et + | 1,38 | 6,26 | 6,26 | 2,31 | 2,58 | 1,19 |

| MAIS1-2 PN | Piéton | Vélo | EDPm | 2RM | VT | Autre |
|---|---|---|---|---|---|---|
| 0-13 ans | 1,89 | 12,61 | 12,61 | 3,66 | 2,33 | 1,29 |
| 14-19 ans | 2,04 | 8,81 | 8,81 | 2,99 | 3,09 | 1,09 |
| 20-29 ans | 1,94 | 7,57 | 7,57 | 2,86 | 2,89 | 1,40 |
| 30-39 ans | 1,79 | 6,71 | 6,71 | 2,56 | 2,64 | 1,41 |
| 40-49 ans | 1,79 | 6,57 | 6,57 | 2,56 | 2,65 | 1,38 |
| 50-59 ans | 1,70 | 5,88 | 5,88 | 2,47 | 2,54 | 1,32 |
| 60-69 ans | 1,53 | 5,17 | 5,17 | 2,28 | 2,29 | 1,23 |
| 70 ans et + | 1,32 | 4,35 | 4,35 | 2,31 | 2,21 | 1,08 |

| MAIS3+ GN | Piéton | Vélo | EDPm | 2RM | VT | Autre |
|---|---|---|---|---|---|---|
| 0-13 ans | 0,39 | 1,09 | 1,09 | 0,89 | 0,22 | 0,20 |
| 14-19 ans | 0,39 | 1,11 | 1,11 | 0,71 | 0,34 | 0,26 |
| 20-29 ans | 0,39 | 1,04 | 1,04 | 0,62 | 0,30 | 0,27 |
| 30-39 ans | 0,41 | 1,19 | 1,19 | 0,59 | 0,29 | 0,27 |
| 40-49 ans | 0,41 | 1,24 | 1,24 | 0,61 | 0,28 | 0,26 |
| 50-59 ans | 0,42 | 1,19 | 1,19 | 0,60 | 0,29 | 0,29 |
| 60-69 ans | 0,46 | 1,15 | 1,15 | 0,69 | 0,31 | 0,28 |
| 70 ans et + | 0,56 | 1,37 | 1,37 | 0,78 | 0,43 | 0,36 |

| MAIS3+ PN | Piéton | Vélo | EDPm | 2RM | VT | Autre |
|---|---|---|---|---|---|---|
| 0-13 ans | 0,15 | 0,42 | 0,42 | 0,19 | 0,04 | 0,03 |
| 14-19 ans | 0,15 | 0,27 | 0,27 | 0,17 | 0,06 | 0,04 |
| 20-29 ans | 0,14 | 0,20 | 0,20 | 0,16 | 0,05 | 0,04 |
| 30-39 ans | 0,15 | 0,22 | 0,22 | 0,16 | 0,05 | 0,05 |
| 40-49 ans | 0,16 | 0,25 | 0,25 | 0,17 | 0,05 | 0,05 |
| 50-59 ans | 0,18 | 0,27 | 0,27 | 0,19 | 0,05 | 0,05 |
| 60-69 ans | 0,21 | 0,35 | 0,35 | 0,22 | 0,06 | 0,05 |
| 70 ans et + | 0,32 | 0,50 | 0,50 | 0,31 | 0,11 | 0,06 |

Lecture : pour la catégorie des piétons de 00-13 ans en zone gendarmerie, le nombre de blessés légers ou modérés MAIS1-2 estimé par l'Univ. Eiffel est 2,82 fois plus important que le nombre total de blessés BAAC (toutes gravités confondues). Le nombre de blessés graves MAIS3+ de cette catégorie est 0,39 fois moins important que le nombre total de blessés BAAC (toutes gravités confondues).

La sécurité routière en France - bilan de l'année 2021 - **ONISR 2022**

Les données d'accidentalité du fichier BAAC proviennent des forces de l'ordre. Si elles sont exhaustives concernant les tués, elles comportent un sous-enregistrement des blessés.

L'Université Gustave Eiffel, en comparant les données du Registre du Rhône[2] et les BAAC, réalise une estimation au niveau national du nombre de blessés afin de donner l'ordre de grandeur de la morbidité routière. Cependant, du fait de la complexité de la méthode et du décalage temporel dans la saisie des données du Registre, l'estimation ne permet pas de suivre en temps réel les évolutions en termes de nombre de blessés. Une méthode simplifiée basée sur les écarts entre les résultats des estimations de l'Univ. Eiffel et les fichiers BAAC a été construite par l'ONISR afin de produire des estimations provisoires concernant les blessés sur les années récentes, en attendant l'estimation plus précise de l'Univ. Eiffel.

### Principe de l'estimation

Le sous-enregistrement des blessés diffère fortement selon le type de l'accident, le milieu routier et les forces de l'ordre : police nationale (PN) ou gendarmerie nationale (GN). Il est par exemple très élevé pour les accidents de cyclistes hors agglomération sans autre tiers impliqué et assez faible pour les accidents graves impliquant deux véhicules motorisés en milieu urbain.

Du fait de la coexistence des deux sources de données dans le département du Rhône (BAAC et Registre) et grâce à une méthode de capture-recapture, l'Univ. Eiffel peut faire une estimation nationale du nombre de blessés. La dernière estimation s'arrête provisoirement à 2016. En comparant les résultats sur 2012-2016 déclinés selon les modes de déplacement, gravité, force de l'ordre et âge de l'usager avec les résultats des BAAC sur la même période, on obtient des ratios entre les blessés BAAC et l'estimation de l'Univ. Eiffel des MAIS1-2 et MAIS3+.

A l'aide de ces ratios calculés sur 2012-2016, on peut multiplier les nombres de blessés BAAC pour chaque année de 2017 à 2021 et obtenir une estimation des blessés déclinée selon le mode, l'âge, la gravité et les forces de l'ordre. Pour les EDPm, peu utilisés avant 2018, l'Univ. Eiffel a identifié sur 2019 des niveaux de sous-enregistrement comparables aux vélos.

En dernière étape, on redistribue les blessés estimés selon le genre grâce à la répartition de l'estimation Registre, et selon le milieu sur la base des répartitions observées dans les BAAC.

En 2017, la mise en place du logiciel Pulsar BAAC en gendarmerie a permis un sous enregistrement moindre de certains types d'accident. Les ratios calculés pour la gendarmerie ont donc été adaptés pour 2017-2021.

[1] Le détail de la méthode complète est disponible sur le site internet de l'ONISR : https://www.onisr.securite-routiere.gouv.fr/
[2] Registre des victimes des accidents de la route d'après les sources hospitalières (voir pages 14 à 16).

Back to section Road Traffic Accident Surveillance

# Appendix I Form to collect core minimum and optional data on any case of Injury (From WHO guidelines[5])

## FORM TO COLLECT CORE MINIMUM AND OPTIONAL DATA ON ANY CASE OF INJURY

| Registration or Identification Number | Date d d m m y y y y | Time h h m m |
|---|---|---|

| Age | Residence |
|---|---|
| Sex | Male / Female / Unknown |

| Place : *Where were you when you were injured?* | | |
|---|---|---|
| 1. Home | 2. School | 3. Highway/Street |
| 8. Other (specify) | | 9. Unknown |

| Activity : *What were you doing when you were injured?* | | |
|---|---|---|
| 1. Work | 2. Education | 3. Sport |
| 4. Travelling | 8. Other (specify) | 9. Unknown |

| Mechanism : *How were you hurt? Or How was the injury inflicted?* | | |
|---|---|---|
| 1. Traffic injury | 2. Sexual Assault | 3. Fall |
| 4. Other Blunt Force | 5. Stab/Cut | 6. Gun Shot |
| 7. Fire, heat | 8. Choking/hanging | 9. Drowning |
| 10. Poisoning | 98. Other (specify) | 99. Unknown |

| Intent | | |
|---|---|---|
| 1. Unintentional | 2. Self-Harm | 3. Intentional (assault) |
| 8. Other (specify) | | 9. Unknown |

| Alcohol Use : *Did you use alcohol within 6 hours of the incident?* | |
|---|---|
| 1. Suspected by report or confirmation | 2. No information |

Back to section Injury surveillance system requirements

# Appendix J TARPON annotation grid

| DIGIT 1 | Première consultation | |
|---|---|---|
| | Oui | 1 |
| | Non | 2 |
| | Incertain | 9 |

| DIGIT 2-3 | Lieu | |
|---|---|---|
| | Intérieur – Domicile ou RPA | 01 |
| | Intérieur – Publique ou collectif | 02 |
| | Extérieur –Jardin ou espace naturel | 03 |
| | Extérieur – Voie ou espace public | 04 |
| | Non renseigné | 99 |

| DIGIT 4-5 | Activité | |
|---|---|---|
| | Domestique – activités vitales, déplacement, ou autre que bricolage et jardinage | 01 |
| | Domestique - bricolage, jardinage | 02 |
| | Professionnelle hors trajet | 03 |
| | Sport | 04 |
| | Loisir hors sport | 05 |
| | Déplacement hors du domicile | 06 |
| | Etudes Ecole | 07 |
| | Non renseigné | 99 |

| DIGIT 6-7 | Sport | |
|---|---|---|
| | Athlétisme | 01 |
| | Aviron | 02 |
| | Basket-Ball | 03 |
| | Bowling | 04 |
| | Boxe | 05 |
| | Danse | 06 |
| | Canoë-Kayak Voile | 07 |
| | Course Footing Running | 08 |
| | Cyclisme | 09 |
| | Equitation | 10 |
| | Escrime | 11 |
| | Football | 12 |
| | Gymnastique | 13 |
| | Golf | 14 |
| | Haltérophilie | 15 |
| | Handball | 16 |
| | Hockey | 17 |
| | Judo | 18 |
| | Musculation | 19 |
| | Natation | 20 |
| | Patinage | 21 |
| | Randonnée | 23 |
| | Rugby | 24 |
| | Ski | 25 |
| | Squash | 26 |
| | Sports de combat hors boxe et judo | 27 |
| | Surf | 28 |
| | Tennis | 29 |
| | Tennis de table | 30 |
| | Volley-ball | 31 |
| | Escalade | 32 |
| | Futsal | 33 |
| | Skate-Board, Roller | 34 |
| | Chasse | 35 |
| | Autre sport | 80 |
| | Non renseigné | 90 |
| | Non applicable | 99 |

| DIGIT 8 | Sujet sous l'emprise | |
|---|---|---|
| | De l'alcool | 1 |
| | Des stupéfiants | 2 |
| | De l'alcool et des stupéfiants | 3 |
| | Incertain (il existe un doute) | 8 |
| | Aucune de ces situations spécifiée | 9 |

| DIGIT 9 | Notion de malaise pré traumatique | |
|---|---|---|
| | Oui | 1 |
| | Pas de notion de malaise | 9 |
| | Incertain (il existe un doute) | 8 |

| DIGIT 10 | AVP - Eléments de prévention secondaire | |
|---|---|---|
| | Ceinture | 1 |
| | Casque | 2 |
| | Aucun des deux | 3 |
| | Applicable mais non renseigné | 4 |
| | Non applicable | 9 |

| DIGIT 11-12 | AVP- antagoniste | |
|---|---|---|
| | Piéton | 00 |
| | Trottinette électrique | 01 |
| | Autre EDP (Hoverboard, skate, ...) | 02 |
| | Bicyclette | 03 |
| | Deux-roues motorisé | 04 |
| | Véhicule léger | 05 |
| | Véhicule utilitaire léger | 06 |
| | Poids Lourd | 07 |
| | Transport en commun | 08 |
| | Obstacle fixe | 10 |
| | Animal | 11 |
| | Pas d'antagoniste | 20 |
| | Applicable mais non renseigné | 90 |
| | Non applicable (non AVP) | 99 |

Back to section Labeling strategy

# Appendix K. TARPON annotation grid

| DIGIT 13-14 | **Accident de la voie publique – Mode de déplacement du patient** | |
|---|---|---|
| | Piéton | 00 |
| | Trottinette électrique | 01 |
| | Autre EDP (Hoverboard, skate, …) | 02 |
| | Bicyclette | 03 |
| | Deux-roues motorisé | 04 |
| | Véhicule léger | 05 |
| | Véhicule utilitaire léger | 06 |
| | Poids Lourd | 07 |
| | Transport en commun | 08 |
| | AVP – Mode de déplacement non renseigné | 09 |
| | **Traumatisme involontaire hors AVP** | |
| | Chute de sa propre hauteur | 10 |
| | Chute escalier | 11 |
| | Chute d'une hauteur | 12 |
| | Coupure | 13 |
| | Choc, écrasement, abrasion, contrainte mécanique | 14 |
| | Perforation | 15 |
| | Brûlure | 16 |
| | Traumatisme acoustique | 17 |
| | Blast, onde de choc | |
| | Electrisation | 18 |
| | Fausse route, asphyxie | 19 |
| | Blessure involontaire par arme à feu | 20 |
| | **Noyade** | |
| | Baignoire et autre contenants domestiques | 21 |
| | Piscine privée | 22 |
| | Piscine publique | 23 |
| | Nature eau douce | 24 |
| | Nature eau salée | 25 |
| | **Animal** | |
| | Morsure | 26 |
| | Piqure insecte | 27 |
| | Animal - Non renseigné ou autre | 29 |
| | **Intoxication** | |
| | Alimentaire | 31 |
| | Médicaments hors psychotropes | 32 |
| | Médicaments Psychotropes | 33 |
| | Stupéfiants | 34 |
| | OH | 35 |
| | CO | 36 |
| | Gaz autre que CO | 37 |
| | Chimique sauf supra | 38 |
| | Intoxication - Non renseigné | 39 |
| | **Traumatisme volontaire** | |
| | TS (Tentative de suicide) par intoxication médicamenteuse ou IMV | 40 |
| | TS par intoxication autres substances | 41 |
| | TS par phlébotomie | 42 |
| | TS par pendaison | 43 |
| | TS par armes à feu | 44 |
| | TS par chute d'une hauteur | 45 |
| | TS autres causes | 46 |
| | TS cause non connue | 47 |
| | Automutilation volontaire | 48 |
| | Agression physique conjoint | 49 |
| | Agression physique autre ou non précisé | 50 |
| | Agression sexuelle conjoint | 51 |
| | Agression sexuelle autre ou non précisé | 52 |
| | Agression physique et sexuelle conjoint | 53 |
| | Agression physique et sexuelle autre ou non précisé | 54 |
| | Automutilation par accès de rage | 55 |
| | Maltraitance physique à enfant par parent | 56 |
| | Détresse psychologique hors agression physique et sexuelle, hors maltraitance. | 57 |
| | Accident d'exposition au sang | 60 |
| | **Corps étranger** | |
| | Œil (liquides) | 70 |
| | Œil (solides) | 71 |
| | Oreille (liquides) | 72 |
| | Oreille (solides) | 73 |
| | Nez (liquides) | 74 |
| | Nez (solides) | 75 |
| | Système digestif | 76 |
| | Anus | 77 |
| | Appareil génital | 78 |
| | Autre | 79 |
| | **Douleur ou incapacité liée** | |
| | A un mouvement | 80 |
| | A un port de charge ou un effort | 81 |

Back to section Labeling strategy

## USING NATURAL LANGUAGE PROCESSING TECHNIQUES TO STUDY AND REGULATE EMERGENCY DEPARTMENT FLOWS

### Development and application to the study of trauma risks based on ED venues in Bordeaux.

**Abstract :**

The TARPON (Traitement Automatique des Résumés de Passage aux urgences dans le but de créer un Observatoire National du traumatisme) project aims to demonstrate the feasibility of setting up a French observatory of trauma. Emergency Departments (EDs) generate a large volume of health-related data and approximately one-third of ED visits are the result of trauma. Most of the information contained in electronic health records is in the form of free text format and manual information extraction is time and resource consuming. Artificial Intelligence (AI) and particularly Natural Language Processing (NLP) could optimize this process. NLP has seen a recent breakthrough with the introduction of deep learning and in particular the Transformer architecture. These Large Language Models (LLMs) have reached the state-of-the-art for most NLP tasks and their use for clinical and medical data is promising.

To explore the potential of Transformers for trauma classification (multi-class), we conducted an evaluation using free-text clinical notes from a single large University Hospital (Bordeaux) ED. A total of 69,110 free-text clinical notes generated between 2012 and 2019 were manually annotated, with 22,481 identified as traumas. To compare the performance of traditional machine learning classifiers and Transformer models, we employed different architectures (BERT and GPT-2), varied sizes, pre-training corpora languages and tokenizers (OSCAR, Wiki, and CCNET). Additionally, we investigated the impact of incorporating a pre-training step on a domain-specific corpus. Our findings revealed that bagging algorithms and Light Gradient Boosting exhibited similar results to the lower-performing Transformers. Interestingly, we discovered that larger models did not necessarily translate to better performance, but the choice of pre-training corpora significantly influenced the outcomes. The best results, with an average F1-score of 0.976, were achieved using a GPT-2 architecture with two steps of pre-training utilizing a French corpus then with a domain-specific corpus. These results highlight the potential of Transformers, particularly when an unsupervised pre-training with a domain-specific corpus is performed, in the accurate classification of traumas based on free-text clinical notes.

Our contribution to the TARPON project laid the groundwork for the use of LLMs for processing clinical notes. These models, which are becoming increasingly efficient and powerful, have led to a recent paradigm shift in NLP. Most AI applications currently in use in emergency medicine are based on NLP and automatic speech recognition because of the privileged documentation medium of free or semistructured text or the practitioner-patient interaction. However, these applications lack proper derivation, validation, or impact evaluations that are performed rigorously and independently. Building a trustworthy, safe, and explainable AI requires a holistic approach that encompasses all sociotechnical aspects involved. Human factors such as participatory design and multistakeholder approaches are important for building such AI systems. Inclusiveness begins at the very beginning of the design step, with the inclusion of stakeholders. All possible biases and risks should be identified and documented before any initiation, and they should be monitored continuously.

However, when emergency medicine is concerned with the development of AI applications, several principles mentioned above collide, and trade-offs must be determined. How can we determine the trade-off among interpretability and performance, time, and explainability? How can transparency be ensured when intellectual property is involved? How can liability be determined when AI harms?

To ensure the safety of patients, healthcare professionals and researchers, we need to bring together all the stakeholders involved in the development of such healthcare tools. Legislators, decision-makers, insurers and public authorities have a duty to work together to provide the best possible support for a change that is taking place in spite of them.

**Keywords** : Artificial Intelligence, Natural Language Processing, Transformer, GPT, Emergencies, Traumas, Surveillance, Public Health

**Unité de recherche**

INSERM U1219, équipe AHeaD, 146 rue Léo Saignat, 33076 Bordeaux