

From Clinical Notes to Structured Data for Secondary Use : Extracting and Normalizing Medical Information with Natural Language Processing

Adam Remaki

▶ **To cite this version:**

Adam Remaki. From Clinical Notes to Structured Data for Secondary Use : Extracting and Normalizing Medical Information with Natural Language Processing. Computation and Language [cs.CL]. Sorbonne Université, 2026. English. ⟨NNT : 2026SORUS203⟩. ⟨tel-05881615⟩

HAL Id: tel-05881615

https://theses.hal.science/tel-05881615v1

Submitted on 25 Nov 2026

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.



Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License







Sorbonne Université

École doctorale Pierre Louis de santé publique

Limics

Structurer les notes cliniques pour la réutilisation des données de santé : extraction et normalisation de l'information médicale par traitement automatique des langues

From Clinical Notes to Structured Data for Secondary Use: Extracting and Normalizing Medical Information with Natural Language Processing

Thèse de doctorat de sciences de données

Par Adam REMAKI

Dirigée par Xavier TANNIER et co-dirigée par Christel GÉRARDIN

Présentée et soutenue publiquement le 13 octobre 2026

Devant un jury composé de :

Anne-Sophie	JANNOT	PU-PH	Présidente
Cyril	GROUIN	IR CNRS, HDR	Rapporteur
Vincent	GUIGUE	PU	Rapporteur
Jackie	CHEUNG	Associate Professor	Examinateur
Julien	VIBERT	Docteur	Examinateur

Except where otherwise noted, this work is licensed under
http://creativecommons.org/licenses/by-nc-nd/3.0/

≮ ⊖⁚⧖⁞⊏。.

Acknowledgments (Remerciements)

I would first like to sincerely thank the members of my jury for taking the time to read and evaluate my thesis, and for their valuable comments. I have chosen to write most of these acknowledgments in French, my mother tongue, to express my feelings and gratitude more freely to all those who have supported me throughout my academic journey.

Ce manuscrit vient clore un long parcours académique, jalonné de nombreux établissements : à commencer par l'école **La Plaine** et ses immenses toboggans verts destinés à évacuer les enfants ; le collège **Jean Moulin** et son deuxième plateau où tout était permis ; le lycée **Saint-Just** et sa vue imprenable sur la ville de Lyon, qui a continué d'émouvoir mes amis et moi bien après le lycée ; la classe **prépa du Parc**, située à deux pas du somptueux parc de la Tête d'Or, dont je n'ai que trop peu profité, vissé à ma chaise au CDI ; **CentraleSupélec** et ses studios de musique flambant neufs, dont j'ai bien plus profité que des salles de cours ; l'**Imperial College London** et sa gestion remarquable de la crise sanitaire ; et enfin, **Sorbonne Université**, dont j'ai surtout connu son authentique passerelle en bois perchée sur le campus des Cordeliers. Mais bien plus que les établissements traversés, ce sont avant tout les nombreuses rencontres qui ont marqué ce parcours et façonné la personne que je suis.

Je vais commencer par la fin du parcours et remercier les personnes qui m'ont encadré pour mener ce travail de thèse jusqu'au bout :
Christel, il faut d'abord rendre à César ce qui appartient à César, comme on dit, et cette opportunité de thèse, c'est d'abord grâce à toi qu'elle est née. Ensuite, je tiens évidemment à te remercier pour ton encadrement tout au long de la thèse. J'ai énormément apprécié que tu sois toujours attentive à ce que cette thèse soit le plus bénéfique pour moi. Que ce soit pour me conseiller quand je suis perdu, pour annoter des documents ou pour se raconter des bêtises autour d'un verre, tu es toujours disponible quand on a besoin de toi malgré la récente PME que tu as lancée l'année dernière. Sérieusement, ton engagement pour l'hôpital public est tellement inspirant. J'espère avoir la chance de recroiser ton chemin et de continuer à œuvrer pour notre hôpital public.
Xavier, je dois d'abord te remercier d'avoir accepté de me prendre comme doctorant. J'avais entendu parler de toi par tes anciens doctorants qui m'ont tous fait tes louanges. Et j'ai pu constater pendant ces trois années qu'ils avaient raison. Tu as toujours su te rendre disponible quand j'avais besoin de toi, malgré la densité de ton emploi du temps de directeur du laboratoire. Je t'ai d'ailleurs bien trop souvent sollicité pour des questions administratives, notamment avec mes enseignements dans trois structures différentes chaque année. J'espère que tu ne m'en veux pas trop ! Mais le plus important pour moi, c'est que tu as vraiment été à l'écoute et que tu m'as toujours fait sentir

que ce qui comptait avant tout pour toi, c'était que cette thèse soit aussi bénéfique que possible pour moi. Je te remercie pour la confiance que tu m'as accordée (même lorsque je décide de soumettre pour la quatrième fois le même article à des conférences toujours plus prestigieuses !) et pour la bienveillance dont tu as toujours fait preuve à mon égard, particulièrement dans les moments les plus difficiles pour moi.
Xavier et Christel, je voudrais aussi rapidement ajouter que, de mon point de vue, votre binôme d'encadrement fonctionne particulièrement bien. Évidemment sur les aspects scientifiques et techniques, avec l'expertise médicale de Christel et l'expertise informatique de Xavier, mais aussi sur le plan humain et relationnel, où je trouve que vous vous complétez bien.

Cette thèse n'aurait bien sûr pas été aussi agréable au quotidien sans la **team passerelle** :
Armand, le boss incontesté de la passerelle (détenteur du tampon officiel). Pour bosser sur des challenges BioNLP, refaire le monde ou manger un CROUScous salade de lentilles, tu réponds présent et c'est un plaisir. On a eu nos périodes d'absence (surtout moi), mais c'est toujours plus facile de revenir quand je sais que tu es là. Merci pour ces beaux moments. Je ne sais pas si ça peut t'intéresser, mais je viens de voir passer une offre de data scientist chez Decathlon ?
Rodrigo, my friend, it was a pleasure to spend time with you in the lab. You were always smiling and happy to help with anything. Thank you, my friend. I hope I will see you again and have the opportunity to dance some salsa.
Morgan, le plus fidèle occupant de la passerelle. Merci pour ta bienveillance et tes efforts pour que la vie du labo soit toujours plus agréable.
Chaï, la boss lady de la passerelle. Merci pour ta joie de vivre au quotidien, on ne s'en lasse pas de t'entendre te marrer.
Philé, le rayon de soleil de la passerelle. Toujours partant pour un CROUS, un verre, une soirée et surtout un concert. Merci pour toutes ces recommandations musicales qui rythment encore mes journées, pour ta curiosité et ta bonne humeur qui nous régalent au quotidien. J'espère qu'on pourra assister à des concerts ensemble.
Lamia et **Ibrahim**, on n'a pas été à la passerelle en même temps, mais c'était un plaisir pour moi de venir me détendre de temps en temps à la passerelle et de discuter avec vous. Merci pour votre présence exemplaire et votre bonne humeur.

Sur la dernière année, j'ai quitté la sombre passerelle pour un bureau plus lumineux. Ce "bureau du fond" a vu se former la **team patient SimSim** :
Guillaume, la stagiaire idéale. On s'est rencontrés à l'occasion de ton stage de M2 et, plus tard, tu as rejoint le labo pour une thèse encadrée par la même dream team que moi. C'était vraiment cool de travailler avec toi mais aussi de discuter de tout et rien.
Marine et **Léna**, c'était très cool de partager ce bureau avec vous. Merci pour votre joie de vivre au quotidien. J'espère qu'un jour on arrivera à le boire ce verre.

Merci évidemment à **Daphné**, **Nour**, **Edgar**, **Massoud**, **Samya**, **Alaedine**, **Akram**, **Manon**, **Nathavy**, **Stella**, **Ons**, **Vikram** et à tous les autres chercheurs qui contribuent ou ont contribué à la belle vie du Limics au quotidien. Enfin, je te remercie, **Isabelle**, pour ton aide précieuse dans toutes mes démarches administratives.

I would also like to extend my sincere thanks to the BSC team in Barcelona: **Martin**, **Eulàlia**, **Edu**, **Jan**, **Judith**, **Miguel**, **Salva**, and **Sergi**. Thank you for welcoming me so warmly into your team during my one-month stay and for making me feel like part of it from the very beginning. I am also grateful for our inspiring discussions and the opportunity to collaborate with you.

Je tiens également à exprimer ma gratitude au **DIM AI4IDF**, un programme de la Région Île-de-France dédié à l'intelligence artificielle, qui a intégralement financé ma thèse. J'adresse un remerciement tout particulier à **Khaled**, qui m'a également offert de nombreuses opportunités de présenter et de valoriser mes travaux à l'occasion de différents événements.

Je remercie évidemment toute **l'équipe Data Science de l'AP-HP** et en particulier **Perceval**, **Basile**, **Thomas**, **Simon**, **Alice**, **Sonia**, **Charline**, **Romain** et bien sûr **Ariel** et **Élise** qui ne cessent de contribuer à rendre l'hôpital public meilleur. Je leur dois également des remerciements plus personnels pour m'avoir tant appris (notamment à faire de magnifiques logos qui finissent au cimetière) pendant les deux années où j'ai eu la chance de faire partie de cette dream team. Mais je les remercie aussi pour avoir continué de m'accueillir dans leurs locaux, que ce soit pour échanger, collaborer ou parfois simplement pour le plaisir de se retrouver. J'ai certes appris plein de choses grâce à vous (edsnlp for ever), mais j'ai surtout fait de belles rencontres qui se sont transformées pour certaines en amitiés qui comptent beaucoup pour moi.

Je ne peux remercier l'équipe Data Science sans remercier également les nombreux chercheurs des URC, qui forment de plus en plus une grande famille avec l'équipe Data Science. Merci à toutes et à tous pour votre engagement et pour votre contribution à une recherche en santé plus rigoureuse et éthique. Je tiens à remercier tout particulièrement celles et ceux avec qui j'ai eu la chance de partager de nombreux jeudis, de collaborer et de partager des cafés-chouquettes : **Kankoé**, **Paul**, **Pierre**, **Mathis**, **Francesco**, **Patricia**, **Richard**, **Judith**, **Marie**, **Lucas** et **Léo**.

Ariel, mon jumeau de thèse, tu es là depuis le début de ma vie professionnelle et tu ne m'as jamais quitté. Depuis mon arrivée au sein de l'équipe Data Science en 2021, on ne s'est pas lâchés. Grâce à toi, le travail est toujours plus agréable : les pauses déjeuner maxi kiff Piscine/Gudule, les pauses théâtre d'impro dans la cour de Picpus et autres petits chit-chats pour se raconter nos vies. Bon, c'est vrai que nos chemins professionnels vont bientôt se séparer. Mais aujourd'hui, notre relation a largement dépassé ce cadre professionnel, donc je ne m'inquiète pas pour nous. Merci, mon Arielos, c'est très agréable de savoir qu'on peut toujours compter sur toi.

Élise, Croco, Bob l'éponge, Baleineau (plus pour longtemps). Quand je pense aux rencontres professionnelles qui se sont transformées en véritables amitiés, c'est typiquement à Ariel et à toi que je pense. Je me rends compte aujourd'hui de la chance que j'avais de travailler au quotidien à tes côtés. Chacun derrière son écran, il suffisait très souvent d'un simple regard pour qu'on se comprenne. Je ne pourrais pas énumérer toutes les choses qui te passionnent et dans lesquelles tu t'investis intensément comme tu sais si bien le faire. C'est inspirant de te voir t'engager autant, et toujours passionnant de t'entendre en parler. Merci, Croco, pour ton écoute attentive, tes bons conseils et surtout pour l'amour et la confiance que tu m'accordes depuis qu'on se connaît.

Cette thèse a débuté de manière assez originale, sur le territoire de la Guyane. Un territoire magnifique qui m'a beaucoup appris et a éveillé tant de choses en moi et surtout grâce aux personnes que j'y ai rencontrées et qui sont devenues des ami·e·s qui me sont cher·ère·s.
Pops, on ne s'est pas rencontrés en Guyane mais c'est quand même grâce à toi qu'on a trouvé cette colocation et qu'on a pu y faire de si belles rencontres. Merci pour ça, mais surtout pour cette intensité que tu mets dans chacun des moments qu'on passe ensemble.
Kimmy, les petits yeux de tortue du matin. Merci de m'avoir fait confiance dès notre rencontre. Merci pour tous ces moments qu'on a partagés ensemble. Merci d'avoir toujours donné le maximum pour qu'on vive intensément chaque instant ensemble.
Mae, le fond, la forme. On s'est souvent retrouvés tous les deux à la maison quand tout le monde partait bosser. C'était l'occasion de manigancer toutes les choses drôles qu'on pourrait faire comme construire une maison pour nos chauves-souris. Merci d'avoir cette petite étincelle en toi que je ne saurais pas vraiment nommer. Cette étincelle qui fait qu'on peut se retrouver à inonder le magasin en dessous de chez nous, mais qu'au final, ça reste une histoire drôle et un beau souvenir.
Nathyfa, la boug dangereuse. Merci pour tous ces beaux moments partagés. Celui de la baignade dans la petite cascade au bord de la route vers Régina restera gravé à jamais. Merci aussi d'avoir contribué à éveiller certaines choses en moi, d'avoir semé de petites graines qui grandissent doucement.
Alexane, le meilleur pour la fin petite fleur. Merci d'être la personne si gentille, sensible et drôle que tu es. Merci pour tous ces moments partagés en Guyane et ailleurs (dans nos différentes colocations). Et enfin, merci de ne pas m'avoir trop jugé quand je suis rentré de l'université de Cayenne avec mon T-shirt à l'envers.
Merci aussi à tous les autres petits cœurs pour tous ces moments si simples et agréables : **Cécéw**, **Maudyto**, **Loanita**, **Delf** et **Robin**.

Et puis il y a les recontres qui se sont faites par l'intermédiaire d'Alice et qui se sont révélées des personnes qui comptent énormément pour moi:
Élise et Alex, je ne sais pas comment vous faites pour que ce soit si agréable de passer du temps avec vous. Je ne sais pas comment vous faites pour qu'on se sente autant en confiance, en lâché prise et simplement apaisé. Je vous remercie pour tous ces moments tout doux ensemble et je vous remercie aussi de nous avoir accueilli dans votre cocon quand on en avait besoin. Honnêtement, c'était dur de quitter cette vie de coloc avec vous.
Apo et Nina, merci d'être aussi inspirantes. Merci d'être ces personnes bienveillantes en amitié et au travail. Enfin, merci pour tous ces moments partagés (même si Apo, t'as essayé de me tuer avec un crumble).

Ces trois années de thèse n'auraient pas été les mêmes si elles n'avaient pas été rythmées par l'atelier Latin Jazz : Salsa Picante. Merci à tous les membres du groupe pour votre engagement tout au long de l'année, votre bienveillance et pour toutes ces émotions fortes vécues ensemble pendant les répètes et les concerts. Pâte à quiche ! Merci **Barnett**, **Ali**, **Benoît**, **Dominique**, **Greg**, **Yasmina**, **Jonathan**, **JB**, **Benjamin**, **Vanessa** et bien sûr **Patrick**.

Je suis très reconnaissant des rencontres que j'ai faites en arrivant à l'école Centrale et qui, pour certaines, sont devenues des amitiés précieuses.
Paul, mon ancien coloc et nouveau voisin. Merci pour toutes ces heures de discussion qui m'enrichissent toujours un peu plus. Merci d'être la personne passionnée et passionnante que tu es. Mais surtout, merci pour la confiance que tu m'accordes et pour être quelqu'un sur qui je sais pouvoir toujours compter. Merci **Mathilde** aussi, c'est trop agréable de te voir, j'espère qu'on arrivera à les organiser ces petits dîners de voisins.
Byb, mon bel éphèbe. Merci pour cette amitié dans laquelle je me sens libre d'être pleinement moi-même. Merci aussi pour toutes nos discussions autour de la transmission et de l'enseignement, qui ont renforcé mon envie d'enseigner à mon tour. Mais surtout, merci pour ces innombrables larges tranches de rigolade.
Louis, le "Yes" man. Merci pour tous ces incroyables moments passés ensemble en salle de répèt, en concert, en soirée (avec Joseph le phoque, par exemple) ou autre. Mais surtout, merci pour ta spontanéité rare que j'admire beaucoup et qui nous a offert tant de bons moments ensemble. Merci à toi aussi, **Kim**, pour ces belles soirées qu'on a partagées ensemble à Paris et Marseille.
Merci à toi, **Pierro**, j'ai hâte de te revoir maintenant que tu vis à Paris.
Merci à l'équipe Pastis Plage pour les petites pensées, les voyages improvisés, les bons moments passés et les expériences sociales. Merci **François**, **Emma**, **Mehdi** et **Youssef**.
Merci aussi à tous les affreux du CE qui régalent de jour comme de nuit, merci mon **Tomingo**, **Yeb**, **Gaël** et **Léoul**.
Merci à Alpaga : **Aurélien**, **Armand**, **Arielle** et **Aziz**, pour tous ces beaux moments de répètes, mais aussi et surtout pour les montées de stress avant les concerts et les émotions pendant et après.

Radim and **Ivicia**, my crazy lovelies. Thank you for those magical moments we share a few times a year, when time seems to stop. With you, I always feel free to be myself. Thank you for everything, and I can't wait for our next time together!

Quand je pense aux rencontres faites au cours de mon parcours qui ont marqué ma vie et façonné la personne que je suis aujourd'hui, c'est évidemment à la team G que je pense. Il faudrait rédiger une autre thèse entière pour énumérer toutes les choses qu'on a faites ensemble et tous les souvenirs que j'ai avec vous.
Mia, Sherlock, Columbo, je ne sais jamais qui est qui. Merci pour cette amitié inconditionnelle. Merci pour cette capacité que j'admire tant à prendre soin des gens que tu aimes. Merci pour ce grain de folie qui nous occasionne bien des fous rires (Mio le ***do, demi-thug). C'est tellement précieux de t'avoir si près de moi. J'espère que ça durera encore le plus longtemps possible, et je pense même qu'un jour, on partagera la même maison.
Valeuuuh, el cabinone, j'ai jamais vraiment su ce que ça voulait dire mais ça te va bien. Je me rends compte de plus en plus à quel point tu es douée pour trouver les bons mots et me donner des conseils quand je suis un peu perdu. Merci aussi de me faire rire, beaucoup rire. Enfin, je suis aussi énormément reconnaissant pour cette petite communauté de vie qu'on a construite, où l'on partage nos légumes et où l'on peut toujours compter les uns sur les autres pour les petites galères du quotidien (comme nous héberger quand on n'a plus de maison).
François et Romain, vous avez joué un rôle particulier pendant cette thèse. Ce mois

passé avec vous a été une véritable parenthèse enchantée. J'ai adoré vivre avec vous. Merci de m'avoir si naturellement fait une place, j'avais vraiment l'impression qu'on formait une petite famille. Mon **Frances**, motomami. Merci pour ces voyages incroyables qu'on a partagés ensemble sur les traces du flamenco et du rebujito. Merci aussi d'être cette personne sincère et attentive.
Lola, Lol la drôle, il n'est pas dingo ce surnom, mais bon, il a une histoire maintenant, je ne peux plus le changer. Merci pour tous ces beaux moments partagés. Chaque fois qu'on se retrouve tous ensemble à Lille, ce sont des moments remplis d'amour et de bonheur. Merci à toi aussi, **Bapt**, c'est toujours agréable de discuter avec toi et j'admire ta capacité à nous apaiser en toute circonstance, à croire qu'anesthésiste c'est fait pour toi. J'ai trop hâte de rencontrer **Joséphine**.
Lucile, ma Luciole. Merci pour cette joie communicative que tu dégages et pour tous ces beaux souvenirs qu'on a ensemble, dont, tout récemment, ce magnifique mariage que vous nous avez offert avec Théo. Merci à toi aussi, **Théo**, pour ta gentillesse et pour tous ces bons moments partagés.
Élodie, mon Elodio. Merci pour tous ces fous rires à chaque fois qu'on se voit. Merci pour ta gentillesse. Merci pour ton hospitalité : je ne compte plus le nombre de vacances de rêve que j'ai passées avec toi et souvent ta famille, en Bretagne ou à la montagne. Et visiblement, cette hospitalité, c'est de famille : on se sent toujours comme un membre de la famille dans ces moments-là, et c'est tellement précieux. J'adore aussi cette façon que vous avez, avec Alex, de profiter de la vie à fond. C'est vraiment inspirant de vous voir ! Merci à toi aussi, **Alex**, de prendre le temps de tant nous transmettre sur tes passions.
Gaët, la paire. Merci pour cette amitié dans laquelle je me sens si bien, complètement moi-même. Je crois qu'on pourrait nous enfermer dix heures dans une pièce sans qu'on voie le temps passer, tellement on trouve toujours des choses à se raconter. Merci pour ta douceur et la confiance que tu me portes. Mais surtout, merci pour tous ces souvenirs incroyables, dont la liste ne cessera de s'allonger, j'en suis sûr. Merci à toi aussi **Estelle**, c'était rare mais à chaque fois qu'on s'est vu c'était vraiment un plaisir !
Georgi, mon Gorg, ma star. Merci pour ta fidélité en amitié et pour cette énergie incroyable que tu dégages à chaque fois qu'on se voit. Merci pour ces discussions sincères, j'admire beaucoup ta franchise. Même si parfois ce n'est pas facile à entendre, je sais qu'elle nous fait toujours avancer. Merci pour tous ces magnifiques souvenirs. Il fallait le faire : réunir toute la team G pour un EVG et faire en sorte que tout se passe aussi bien ! Bsartek ! **Maud**, merci d'être cette personne rayonnante et bienveillante en toute cironcstance, c'est tellement inspirant. Merci aussi pour ces soirées créatives ! Tu ne peux pas savoir comme ça m'a vidé la tête à des moments où j'en avais besoin (même si mon cadre en pâte autodurcissante était trop fin et a fini par se briser) !
Benjo, mon test'. Merci d'être cet ami sur qui je sais que je pourrai toujours compter. Cette confiance mutuelle a commencé à se construire solidement au fil des fameuses discussions post-volley (bonjour les enfants), et elle ne fait que grandir avec les années. Merci pour ta fidélité sans faille en amitié. Merci enfin de ne jamais m'en vouloir quand j'arrive en retard ou quand je te dis que je te rappelle dans cinq minutes et que je ne le fais que deux heures plus tard, voire plus. J'essaie de m'améliorer, promis, même si parfois je replonge ! Merci d'être toujours patient avec moi, je t'aime trop. Merci à toi aussi, **Loulou**, déjà pour tous ces week-ends où vous m'avez accueilli, même si vous me rackettez des sacs de couchage en échange. Mais surtout, merci pour ta présence, c'est tellement agréable de passer du temps avec toi.

Luca, ma crapule. Mon binôme inséparable (jusqu'en 2100 poto). Merci pour tous ces souvenirs, pour ces riches et nombreux papotages dont je ne me lasserai jamais, pour les aventures intenses qui nous ont parfois fait frôler la mort. On a beau s'éloigner géographiquement, la vie finit toujours par nous rapprocher. Après six mois passés sur deux continents différents, on s'est retrouvés à vivre sous le même toit à Cayenne. Alors je ne m'inquiète pas trop, je suis sûr que la vie trouvera bientôt un moyen de nous réunir à nouveau dans la même ville.
Vic, mon cou, mon acolyte maléfique des 400 coups. Merci tout simplement pour cette amitié unique et si précieuse qui m'accompagne depuis plus de la moitié de ma vie maintenant. Merci pour tous ces fous rires qu'on a ensemble pour tout et rien, mais surtout rien (le problème avec les fruits, c'est qu'ils sont hyper tensés). Je ne m'en lasserai jamais. Des fois, j'aimerais pouvoir faire comme Sergent Cortes et prendre un tunnel pour revivre des moments du passé ou bien rencontrer mon acolyte du futur. C'est le dernier remerciement que j'écris, les mollets sont crispés, les paupières sont lourdes et je commence à écrire n'importe quoi. Alors je vais finir simplement : merci pour tout ça et je te kiffe grave.

Je ne peux pas remercier la team G sans remercier aussi mes voisins et meilleurs amis d'enfance, **Gilles** et **Alex**. Même si, avec Alex, on s'est un peu perdus de vue, je serai toujours reconnaissant de cette enfance incroyable qu'on a partagée tous les trois, à passer nos journées fourrés ensemble. Et mon Gilou, je suis trop reconnaissant qu'on ait réussi à garder ce lien malgré le temps qui passe.

Enfin, j'aimerais remercier ma famille:
Rirouche, le diamant d'innocence. Merci pour la bienveillance sans faille dont tu as toujours fait preuve à mon égard. Merci aussi pour toutes ces discussions si riches que nous avons au quotidien. Je ne cesse d'apprendre des choses avec toi, mais je me sens aussi toujours attentivement écouté. Ces dernières années ont été particulièrement fortes pour nous, même si ce n'était pas toujours facile, j'ai le sentiment que cela n'a fait qu'ouvrir encore plus profondément la communication entre nous.
Nanou, comédienne, conteuse, lectrice à voix haute, mais surtout ma grande sœur adorée. Heureusement que tu es là pour poser les vraies questions et entrer profondément dans des contrées dans lesquelles personne n'ose s'aventurer. Merci d'avoir toujours été attentive à mon bien-être et d'avoir toujours veillé à cultiver ma sensibilité.
Etienne, le petit prince du scratch, merci pour tout ce que tu as fait pour nous, en particulier ces dernières années.
Mara, ma nièce adorée, la grâce incarnée, tu ne cesses de nous épater. Merci pour cette étincelle de bonheur que tu fais rayonner en permanence autour de toi. J'ai hâte que tu viennes à Paris.
Rosalie-Jeanne-Gustaaaaave, les petits anges tombés du ciel, j'ai hâte de continuer de vous voir grandir. Merci d'apporter cette graine d'insouciance et de peps à chaque fois que je vous vois.
Maman, plus le temps passe, plus je me rends compte de la chance que j'ai d'avoir une mère comme toi. Depuis ma naissance, tu m'as offert un cadre de vie tellement sécurisant, entouré de ta bienveillance et de ton amour inconditionnel. Je ne pourrai jamais assez te remercier pour tout cet amour que tu m'as donné et que tu me donnes encore.

Alice, la boss lady, la queen. On va bientôt pouvoir dire qu'on a passé la moitié de notre vie ensemble. Ce serait donc bien trop long d'énumérer toutes les choses pour lesquelles je te suis reconnaissant. Toutes ces petites et grandes choses que tu fais pour moi au quotidien, toutes les valeurs auxquelles tu crois, que tu défends parfois avec conviction et qui m'inspirent moi aussi, et surtout tout l'amour et le soutien que tu me donnes et qui me portent chaque jour. Merci pour tout ça, et pour tant d'autres choses encore. Tu es ma personne préférée, je t'aime.
Frédérique, l'infatigable traileuse. Merci pour tous ces souvenirs partagés ensemble. Votre énergie est vraiment inspirante. Mais surtout, merci pour votre bienveillance et pour tout le soutien que vous nous avez apporté, à ma famille et à moi, particulièrement dans les moments où nous en avions le plus besoin.
Je remercie profondément aussi ma famille maternelle avec qui nous avons traversé des moments particulièrement difficiles cette année. Ça a aussi été l'occasion de partager des moments inoubliables tous ensemble en Algérie pour rendre hommage à notre Mémé adorée. Merci ma **Tata Jacqueline**, **Thierry**, **Eylann**, **Jimmy**, **Nadir**, **Iryll**, **Clémentine**, les petits gremlins **Celeste** et **SoSo** et bien sûr ma petite **Méméy**.
Je remercie aussi mes oncles, tantes, cousins et cousines paternels pour votre immense générosité, votre gentillesse et tous ces moments incroyables qu'on a passés ensemble cet été. Mais surtout, je vous remercie pour tout le soutien et l'amour inconditionnel que vous m'avez apportés et continuez à m'apporter, particulièrement quand j'en ai eu le plus besoin. Merci **Tata Kouka**, **Tonton Mahfoud**, **Tata Djo**, **Tonton Rafik**, **Tata Zira**, **Tonton Karim**, **Tonton Jeff**, **Tata Karima**, **Tonton Lakhdar**, **Tata Malika**, **Tonton Omar** et **Tata Zahra**, **Kahina**, **Dadoune**, **Larbi**, **Kamo**, **Mila**, **Monssaf**, **Amayes**, **Océane**, **Rémi**, **Sarah**, **Lyes**, **Sarah** et tous les autres que je n'ai pas encore rencontrer.

Bon, maintenant j'aimerais profiter de ces remerciements pour vous parler d'un figuier. Pendant cette thèse, la feuille d'un figuier est tombée. Cette grande feuille protégeait plein d'autres feuilles du soleil et des insectes. Après sa chute, les feuilles de la même branche se sont retrouvées sous les rayons brûlants du soleil caniculaire, à la merci des insectes. C'était le 7 juin 2025, il faisait chaud, les feuilles étaient immobiles et penchaient vers le sol, tirées par la gravité de la situation. À ce moment-là, quelque chose de presque surnaturel s'est produit. Toutes les feuilles des autres branches du figuier se sont rapprochées. Certaines se sont mises dessous pour soutenir les feuilles tombantes, d'autres se sont mises au-dessus pour les protéger du soleil et d'autres encore gesticulaient dans tous les sens pour faire fuir les insectes. Merci du plus profond de mon cœur à toutes les feuilles du figuier, celles qui sont tombées comme celles qui sont encore accrochées à ses branches.

Papa, la grande feuille du figuier, je sais que tu n'aurais pas voulu que j'en fasse trop, et qui plus est dans mon manuscrit de thèse qui pourra être consulté par n'importe qui. Donc je vais limiter la voilure, mais bon, il faut quand même te rappeler que c'est toi qui es à l'origine de cette thèse, c'est toi qui me pousses depuis la fin de mes études à la faire pour devenir enseignant-chercheur. Me voilà au bout, forcé de constater que ça m'a quand même bien plu, donc je te remercie d'avoir planté cette graine et de l'avoir souvent arrosée (presque un peu trop parfois). Sourire aux lèvres, les mains qui se frottent, c'est la vie de château pourvu que ça dure.

Summary

Abstract

This thesis investigates how Natural Language Processing (NLP) can transform unstructured clinical text into structured variables to support the secondary use of Electronic Health Records (EHRs). EHRs contain large amounts of data, but only part of this information is available in structured and directly queryable tables. Yet, a large part of clinical information remains in clinical notes, which provide richer descriptions of patients but must first be structured before they can be reused at scale.

The thesis addresses this challenge through three complementary contributions. First, it studies an end-to-end NLP pipeline designed to structure information from discharge summaries in the Greater Paris University Hospitals clinical data warehouse. The pipeline extracts laboratory tests and drug treatments, interprets their clinical context, including negation, uncertainty, conditional statements, and temporality, and links them to standard concepts so that they can be combined with already structured EHR data. This work was evaluated in a cohort of 6,891 patients with immune-mediated inflammatory diseases and applied to 22,194 discharge summaries. The pipeline reached an end-to-end F1-score of 71.1 for laboratory test normalization and 89.3 for drug name normalization. At the cohort level, adding discharge summaries identified positive antinuclear antibodies in 53.6% additional patients with systemic lupus erythematosus compared with structured data alone.

This first contribution shows the practical value of structuring clinical text, but also highlights a key limitation: extracted mentions must be reliably linked to standard concepts. The thesis therefore focuses next on biomedical entity linking (BEL), which maps textual mentions to concepts in a knowledge base (KB). In the first pipeline, BEL mainly relied on the mention itself, which can fail when the same surface form is ambiguous and the correct concept depends on context. Context-aware BEL models can address this problem, but they require costly expert annotations that cover only a small fraction of the target KB. To reduce this bottleneck, the thesis introduces SynCABEL, a synthetic contextualized augmentation framework. SynCABEL uses large language models to generate contextual examples from KB concepts and combines them with human annotations to train generative BEL models. Experiments on multilingual benchmarks

show that SynCABEL improves BEL, reaching an average Recall@1 of 75.4 with Llama-3-8B and guided inference, and matching full human supervision with up to 60% fewer annotations.

Third, the thesis studies document-level consistency in BEL. Most approaches link each mention independently, often using only the local sentence, even though biomedical documents frequently contain repeated concepts, abbreviations, and complementary cues. To address this limitation, the thesis introduces LongBEL, a document-level generative framework that conditions each prediction on the full document and on a memory of previous linking decisions. This allows the model to reuse information introduced earlier in the document when linking later mentions. Across three datasets in English, French, and Spanish, LongBEL outperformed the sentence-level model. The gains were largest for concepts already mentioned in the document and when such recurrences were sufficiently represented in the training data.

Overall, this thesis shows that NLP can transform part of the information contained in free-text clinical notes into structured variables that complement existing EHR data for secondary use. It also addresses two key limitations of BEL: annotation scarcity and the lack of consistency across predictions within the same document. Finally, the reuse of the end-to-end pipeline across at least seven clinical research projects illustrates its practical value beyond the methodological evaluations presented in this thesis.

Keywords: Natural Language Processing, Biomedical, Information Extraction, Entity Linking, Electronic Health Record, Clinical Data Warehouse

Résumé en Français :

Cette thèse étudie comment le traitement automatique des langues (TAL) peut transformer les textes cliniques en variables structurées pour faciliter l'usage secondaire des dossiers patients informatisés (DPI). Ces dossiers contiennent de grands volumes de données, mais seule une partie est structurée et directement requêtable. Or, une large part de l'information clinique se trouve dans les notes, qui décrivent plus finement les patients mais doivent être structurées avant d'être réutilisées à grande échelle.

La thèse répond à ce défi à travers trois contributions. Premièrement, elle étudie une chaîne complète de TAL appliquée aux comptes rendus d'hospitalisation de l'entrepôt de données de santé de l'Assistance Publique – Hôpitaux de Paris. Cette chaîne extrait les tests biologiques et les traitements, interprète leur contexte clinique, notamment la négation, l'incertitude et la temporalité, puis les associe à des concepts standardisés afin de les combiner avec les données déjà structurées. Ce travail a été évalué sur une cohorte de 6 891 patients atteints de maladies auto-immunes ou inflammatoires systémiques et appliqué à 22 194 comptes rendus. Cette approche a atteint un F1-score de 71,1 pour les tests biologiques et de 89,3 pour les traitements. À l'échelle de la cohorte, l'ajout des comptes rendus a permis d'identifier des anticorps antinucléaires positifs chez 53,6 % de patients supplémentaires atteints de lupus par rapport aux données structurées seules.

Dans un second temps, la thèse se concentre sur la normalisation, qui associe les mentions textuelles aux concepts correspondants dans une base de connaissances. Cette étape est essentielle pour rendre les informations extraites analysables à grande échelle. Dans la première contribution, elle reposait sur la mention elle-même, ce qui peut échouer lorsque le concept correct dépend du contexte. Les modèles contextuels peuvent répondre à ce problème, mais nécessitent des annotations expertes coûteuses. La thèse introduit donc SynCABEL, une méthode d'augmentation synthétique contextualisée. SynCABEL utilise de grands modèles de langue pour générer des exemples contextualisés à partir de concepts d'une base de connaissances, puis les combine avec des annotations humaines pour entraîner des modèles génératifs de normalisation. Des expériences sur des jeux de données en anglais, français et espagnol montrent que SynCABEL améliore la normalisation biomédicale, avec un Recall@1 moyen de 75,4, tout en réduisant jusqu'à 60 % le besoin en annotations humaines.

Dans un troisième temps, la thèse s'intéresse à la cohérence des prédictions à l'échelle du document. La plupart des approches, même contextuelles, normalisent chaque mention indépendamment, souvent à partir de la phrase seule. Pourtant, les documents biomédicaux contiennent des concepts récurrents ou liés entre eux : un même concept peut être répété sous différentes formes, ou un symptôme se rapporter à une maladie mentionnée plus tôt. Pour dépasser cette limite, la thèse introduit LongBEL, une méthode de normalisation générative qui conditionne chaque prédiction sur le document complet et sur les prédictions précédentes. Sur trois jeux de données en anglais, français et espagnol, LongBEL améliore le modèle limité à la phrase. Les gains concernent surtout les concepts déjà mentionnés dans le document et sont plus importants lorsque ces récurrences sont suffisamment représentées dans les données d'entraînement.

Dans l'ensemble, cette thèse montre que le TAL peut transformer une partie des textes cliniques en variables structurées qui complètent les données déjà disponibles dans les DPI. Elle traite aussi deux limites de la normalisation : la rareté des annota-

tions et le manque de cohérence entre les prédictions d'un même document. Enfin, la réutilisation de la chaîne complète de TAL dans au moins sept projets cliniques illustre son intérêt pratique au-delà des évaluations méthodologiques présentées dans cette thèse.

Mots-clés : Traitement Automatique des Langues, Biomédical, Extraction d'information, Normalisation, Dossier Patient Informatisé, Entrepôt de Données de Santé

Contents

Summary **xiii**

List of Acronyms and Abbreviations **xxi**

Introduction **1**
1 Context and motivation . . . 1
1.1 Secondary use of Electronic Health Records . . . 1
1.2 Clinical Data Warehouses . . . 1
1.3 EHR data . . . 2
1.4 Natural Language Processing of clinical reports . . . 3
2 Problem Statement and Research Objectives . . . 4
3 Contributions of the thesis . . . 4
4 Outline . . . 5
5 Associated Publications and Preprints . . . 5

1 Background **7**
1 Real-World Data in Healthcare . . . 7
1.1 EHRs as Real-World Data . . . 7
1.2 From Care Delivery to an OMOP Research Database . . . 9
1.3 Biases in CDW Data . . . 11
2 Natural Language Processing for Clinical Text . . . 13
2.1 A Brief History of Natural Language Processing . . . 14
2.2 Biomedical Named Entity Recognition . . . 15
2.3 Clinical Entity Qualification . . . 19
2.4 Biomedical Entity Linking (Entity Normalization) . . . 22
3 Conclusion . . . 27

2 Enriching Phenotyping Through Discharge Summary Processing **29**
1 Introduction . . . 29
1.1 Context . . . 29
1.2 Objectives and Contributions . . . 30
1.3 Related Work . . . 31

2 Methods . . . 33
2.1 Selected Diseases . . . 33
2.2 Dataset Selection . . . 35
2.3 Knowledge Bases . . . 35
2.4 End-to-End Pipeline . . . 37
2.5 Ethical Considerations . . . 42
3 Results . . . 42
3.1 Dataset Description . . . 42
3.2 NLP Performance . . . 42
3.3 Clinical Application . . . 50
4 Discussion . . . 53
4.1 Main Findings . . . 53
4.2 Clinical Relevance . . . 54
4.3 Limitations . . . 55
5 Conclusion . . . 56

3 Synthetic Contextualized Augmentation for Biomedical Entity Linking 57
1 Introduction . . . 57
1.1 Context . . . 57
1.2 Objectives and Contributions . . . 58
1.3 Related Work . . . 59
2 Methods . . . 61
2.1 Problem Statement . . . 61
2.2 Generative BEL Training Objective . . . 61
2.3 Adaptive Concept Representation for Training . . . 62
2.4 Synthetic Data Generation . . . 63
2.5 Training Data Composition . . . 63
2.6 Guided Inference . . . 63
2.7 LLM-as-a-judge Evaluation . . . 64
3 Experiments . . . 64
3.1 Datasets . . . 64
3.2 Implementation Details . . . 67
3.3 Results . . . 67
4 Discussion . . . 71
4.1 Annotation Scarcity . . . 71
4.2 Reducing Annotation Effort . . . 72
4.3 Practical Considerations: Precision and Inference Footprint . . . 73
4.4 End-to-End Impact of Context-Aware Normalization . . . 75
4.5 Limitations . . . 76
5 Conclusion . . . 78

4 Long-Context and Document-Consistent Biomedical Entity Linking 81
1 Introduction . . . 81
1.1 Context . . . 81
1.2 Objectives and Contributions . . . 82
1.3 Related Work . . . 83
2 Methods . . . 84

2.1 Document-Level Autoregressive Formulation . . . 84
2.2 Training . . . 85
2.3 Inference . . . 85
3 Experiments . . . 87
3.1 Datasets . . . 87
3.2 Implementation Details . . . 88
3.3 Results . . . 88
4 Discussion . . . 91
4.1 Impact of Concept Redundancy . . . 91
4.2 Detailed Subset Analysis . . . 92
4.3 Qualitative Examples . . . 92
4.4 Robust Memory Mitigates Cascading Errors . . . 98
4.5 Generalization Across Datasets . . . 98
4.6 Saliency Maps Illustrate the Use of Document-Level Evidence . . . 99
4.7 Efficiency Trade-offs . . . 100
4.8 End-to-End Impact of Document-Level Normalization . . . 102
4.9 Limitations . . . 103
5 Conclusion . . . 104

General Conclusion **105**
1 Answer to the Research Question . . . 105
2 Answer to the Research Objectives . . . 107
2.1 Objective 1: Transforming Clinical Notes into Structured Variables 107
2.2 Objective 2: Reducing the Annotation Bottleneck in Biomedical Entity Linking . . . 107
2.3 Objective 3: Improving Document-Level Consistency in Biomedical Entity Linking . . . 107
3 Impact on Clinical Research . . . 108
3.1 Reuse of the Pipeline in Clinical Research Projects . . . 108
3.2 Potential Uses for Future Clinical Studies . . . 109
4 Discussion and Perspectives . . . 111
4.1 Limits of Predefined Structured Extraction . . . 111
4.2 Enriching Structured Variables with Relations . . . 112
4.3 Other Paradigms for Accessing Clinical Information . . . 112
4.4 LLM Agents and Structured Extraction as Complementary Approaches 113
4.5 Perspectives for Biomedical Entity Linking . . . 113
4.6 Overall Perspective . . . 114

Extended French summary – Résumé étendu en français **115**
1 Introduction . . . 115
1.1 Contexte . . . 115
1.2 Problématiques et objectifs . . . 117
2 Une chaîne de TAL pour transformer les comptes rendus en données structurées . . . 118
2.1 Matériel et méthodes . . . 118
2.2 Résultats et discussion . . . 120
2.3 Conclusion . . . 121

3 Compléter des annotations par des données synthétiques pour la normalisation . . . 121
3.1 Matériel et méthodes . . . 121
3.2 Matériel et méthodes . . . 122
3.3 Résultats et discussion . . . 123
3.4 Conclusion . . . 125
4 LongBEL : normalisation biomédicale cohérente à l'échelle du document . 125
4.1 Matériel et méthodes . . . 125
4.2 Résultats et discussion . . . 127
4.3 Conclusion . . . 128
5 Synthèse et perspectives . . . 129

References **131**

Appendices **147**
A Supplementary Materials for Chapter 2 . . . 147
A.1 ICD-10 Codes for Diagnoses . . . 147
A.2 Annotation Guidelines . . . 148
A.3 NER – Input Labels Used for GLiNER-based NER . . . 150
A.4 NER – Prompt Used for LLM-based NER . . . 151
A.5 NER – Comparison of Different Pretrained BERT Models . . . 152
A.6 NER – Performance on the Quaero Dataset . . . 153
A.7 Standard Codes for Studied Tests and Drugs . . . 154
B Supplementary Materials for Chapter 3 . . . 156
B.1 Preliminary Comparison of Encoder–Decoder Backbones . . . 156
B.2 Prompt Examples . . . 157
B.3 NER Performance by Entity Type . . . 160
B.4 End-to-End Threshold Selection Curves . . . 160
B.5 Preliminary Comparison of Human and Random Concept Distributions . . . 161
C Supplementary Materials for Chapter 4 . . . 164
C.1 Semantic-Group Comparison with SynCABEL . . . 164
C.2 Fine-grained Analysis by Semantic Group . . . 164
C.3 Robustness to Noisy Memory . . . 167
C.4 End-to-End Threshold Selection Curves . . . 167

List of Figures **169**

List of Tables **175**

List of Acronyms and Abbreviations

API	Application Programming Interface
APS	Antiphospholipid Syndrome
ATC	Anatomical Therapeutic Chemical
BEL	Biomedical Entity Linking
BioNER	Biomedical Named Entity Recognition
CCAM	*Classification commune des actes médicaux*
CDW	Clinical Data Warehouse
CI	Confidence Interval
CNIL	*Commission nationale de l'informatique et des libertés*
CNN	Convolutional Neural Network
CRF	Conditional Random Field
CRP	C-Reactive Protein
CUI	Concept Unique Identifier
CWME	Copy Wrong Memory Error Rate
DPI	*Dossiers patients informatisés*
EDS	*Entrepôt de Données de Santé*
EHR	Electronic Health Record
ETL	Extract, Transform, and Load
GFR	Glomerular Filtration Rate
ICD-10	International Classification of Diseases, Tenth Revision
IE	Information Extraction
I2B2	Informatics for Integrating Biology and the Bedside
INR	International Normalized Ratio
JW	Jaro–Winkler
KB	Knowledge Base
LEV	Levenshtein
LLM	Large Language Model
MCP	Model Context Protocol
MeSH	Medical Subject Headings
MM-ST21pv	MedMentions-ST21pv
MMD	Maximum Mean Discrepancy

NER	Named Entity Recognition
NLP	Natural Language Processing
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
RRF	Reciprocal Rank Fusion
RWD	Real World Data
SLE	Systemic Lupus Erythematosus
SNOMED CT	Systematized Nomenclature of Medicine – Clinical Terms
SPACCC	Spanish Clinical Case Corpus
TA	Takayasu Arteritis
TAL	*Traitement automatique des langues*
UMLS	Unified Medical Language System

Introduction

This manuscript investigates how Natural Language Processing (NLP) methods can be used to transform unstructured clinical data into structured variables that support the secondary use of electronic health records.

1 Context and motivation

1.1 Secondary use of Electronic Health Records

The rapid digitization of healthcare has profoundly transformed clinical practice, leading to the large-scale accumulation of routinely collected patient data. Typically captured during care delivery, this information is stored in the *electronic health record* (EHR). According to the International Organization for Standardization, an EHR is a secure digital repository of patient data that can be stored, exchanged, and accessed by multiple authorized users (2005). While originally designed to support patient care, EHRs have increasingly become a crucial resource for secondary uses, including observational studies, healthcare quality improvement, and the development of data-driven decision support tools. Moreover, their large scale and longitudinal nature make EHR data particularly well suited for retrospective cohort studies, which rely on pre-existing records and therefore substantially reduce both the cost and time required for data collection (Casey et al., 2016).

1.2 Clinical Data Warehouses

To make EHR data accessible for secondary use, many institutions have developed clinical data warehouses (CDWs), which integrate EHR data while providing secure access. An early and influential example is the STARR platform (formerly known as STRIDE) at Stanford (Lowe et al., 2009). In France, a prominent example is provided by the Greater Paris University Hospitals (Assistance Publique–Hôpitaux de Paris, AP-HP), the largest hospital network in Europe and one of the largest worldwide. It comprises 38 hospitals, employs more than 100,000 healthcare professionals, and treats approximately 8 million patients each year. Since 2012, the progressive deployment of a shared EHR software across AP-HP hospitals has led to the continuous accumulation of large

volumes of EHR data. Building on this infrastructure, AP-HP has developed a dedicated CDW, known as the *Entrepôt de Données de Santé* (EDS), which aggregates EHR data from all 38 hospitals and supports secondary uses of EHR data, including large-scale research and operational applications. The secondary use of health data within the EDS was authorized by the French Data Protection Authority, the *Commission nationale de l'informatique et des libertés* (CNIL), in 2017 (Commission nationale de l'informatique et des libertés, 2017).

1.3 EHR data

EHR data availabe in CDWs are typically divided into two main categories: structured and unstructured data. Structured data correspond to standardized, directly queryable elements, such as laboratory test results or diagnostic codes from the International Classification of Diseases, Tenth Revision (ICD-10). In contrast, unstructured data include free-text clinical notes and other non-tabular modalities such as medical imaging.

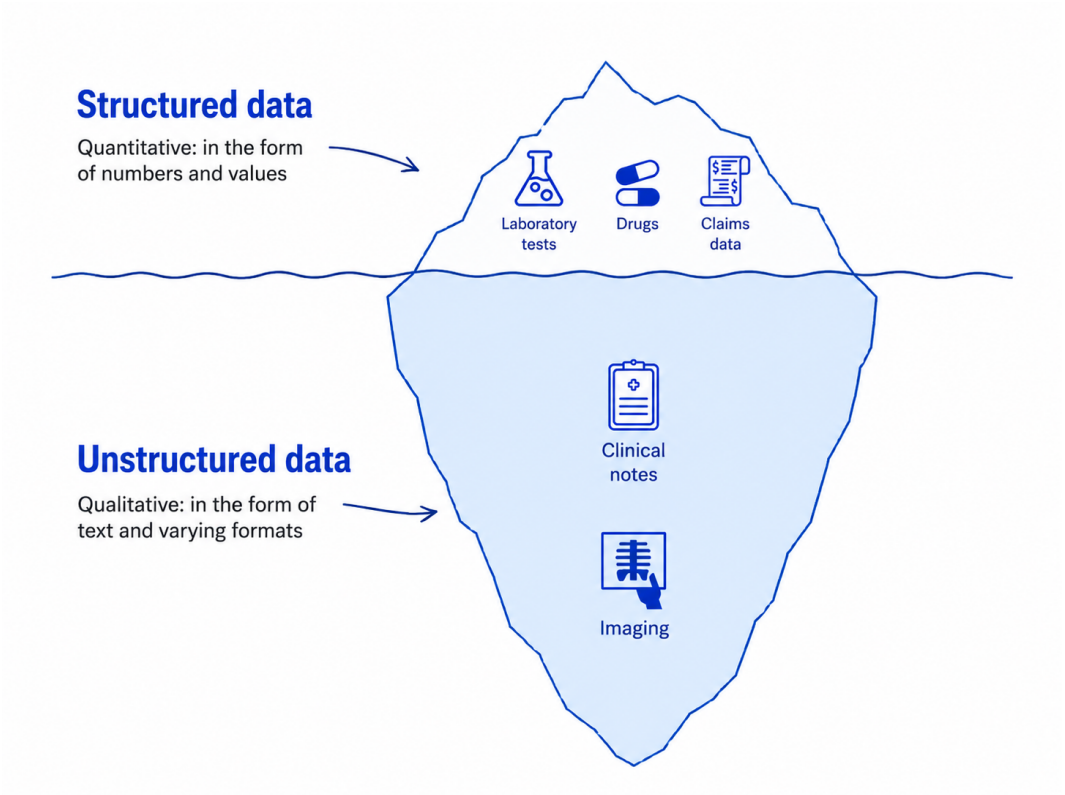


Figure 1: Iceberg representation of EHR data: structured data form the visible and easily accessible portion, whereas unstructured data form the larger, submerged portion, which contains richer information but is more difficult to access.

While structured data are readily usable for large-scale analyses, they often provide an incomplete view of the patient's clinical history. In particular, data collected in CDWs are frequently limited to intrahospital records and may not capture care delivered in external settings. For example, laboratory tests performed prior to hospitalization or in external laboratories may be missing. This limitation is especially problematic for lon-

gitudinal analyses, where access to a patient's full clinical trajectory is essential. More fundamentally, structured data capture only a subset of clinical information, restricted to predefined variables and standardized coding systems. In contrast, unstructured clinical notes provide a richer and more nuanced description of the patient's condition, including medical history, symptoms, treatment trajectories, physician reasoning, and contextual details that are not present in the structured data. As a result, a substantial portion of clinically relevant information is only available in free-text form.

As illustrated in Figure 1, EHR data can be viewed as an iceberg: structured data form only the visible and easily accessible portion, while the majority of clinically valuable information lies beneath the surface in unstructured data. Although this hidden layer is richer, it is also significantly more challenging to extract and analyze. A key challenge for the secondary use of EHR data is therefore the automatic extraction of information from clinical notes with NLP.

1.4 Natural Language Processing of clinical reports

The application of NLP to clinical reports has emerged as a key approach to unlock the vast amount of information contained in unstructured text within CDWs (Bazoge, 2021). These textual data, although rich in clinical detail, cannot be manually exploited for large-scale analysis, thereby motivating the use of NLP techniques. In this thesis, we focus on *information extraction* (IE), which aims to transform free-text into structured, machine-readable data. This transformation is essential for the secondary use of EHRs, supporting research activities such as retrospective studies, cohort selection, and clinical decision support, as well as hospital management.

IE in clinical NLP is typically modeled as a pipeline of interdependent subtasks. The first step is **Named Entity Recognition** (NER), which consists of detecting spans of text corresponding to clinically relevant entities and assigning them a semantic type, such as disease, medication, procedure, or laboratory measurement. In other words, NER determines both the boundaries of an entity mention, from its first to its last character in the text, and the category to which this mention belongs. This step provides the foundational elements upon which subsequent processing relies. However, identifying and typing entity spans alone is insufficient to capture their full meaning. Clinical narratives require additional interpretation through **Qualification**, which includes detecting contextual attributes such as negation (e.g., exclusion of a diagnosis), temporality (distinguishing past from current conditions), and levels of certainty. Qualification may also involve more fine-grained attributes, such as identifying whether a treatment is being initiated or discontinued, or interpreting temporal expressions (e.g., absolute dates versus relative references). Building upon these enriched entities, **Relation Extraction** aims to identify semantic links between them. For instance, it enables the association of a treatment with its dosage, or a surgical procedure with its corresponding date. These relations are critical for reconstructing coherent and clinically meaningful information from fragmented textual mentions. A final and essential step is **Biomedical Entity Linking** (BEL), also referred to as entity normalization, which maps textual mentions to standardized concepts in a controlled knowledge base (KB). This process resolves the variability of clinical language by mapping synonyms, abbreviations, and ambiguous expressions to shared concepts. It is crucial for enabling large-scale analyses and for combining structured and unstructured data.

Beyond information extraction, other NLP tasks are required to enable the secondary use of EHRs. In Europe, health data are subject to specific protections under the General Data Protection Regulation, and in France the CNIL provides a dedicated reference framework for CDWs. This framework allows EHR data to be reused under specific safeguards, including access control, traceability, and pseudonymization (European Parliament and Council of the European Union, 2016; Commission nationale de l'informatique et des libertés, 2021). In this context, de-identification plays a central role for clinical notes by automatically detecting and removing sensitive personal information, such as patient names, dates, or locations, while preserving the clinical content (Tannier et al., 2024).

2 Problem Statement and Research Objectives

To what extent can Natural Language Processing methods transform clinical notes into structured data suitable for the secondary use of electronic health records?

To address this question, this thesis pursues three main objectives.

Objective 1: Evaluate the practical value of end-to-end information extraction. This objective examines the performance of an NLP pipeline applied to discharge summaries and assesses whether the extracted information can complement structured EHR data for secondary use.

Objective 2: Reduce the annotation bottleneck in biomedical entity linking. This objective investigates how synthetic contextualized data can reduce the dependence of BEL models on costly expert annotations and improve their coverage of the target knowledge base.

Objective 3: Improve document-level consistency in biomedical entity linking. This objective studies whether full-document context and previous linking decisions can improve the normalization of mentions across a document.

3 Contributions of the thesis

This thesis makes three contributions to the transformation of clinical narratives into structured information for the secondary use of EHR data.

First, we design and evaluate a complete end-to-end information extraction pipeline for discharge summaries within a CDW. This contribution provides a practical foundation for the thesis by showing how information extracted from clinical text can be combined with existing structured EHR data. It also highlights the main limitations of such pipelines in real-world settings.

Second, we address the annotation bottleneck of BEL, a central limitation of the pipeline identified above. We propose SynCABEL, a synthetic data augmentation strategy designed to reduce the dependence of context-aware normalization models on costly expert annotations and to improve their coverage of large biomedical KBs.

Third, we investigate the consistency of BEL within complete documents. We introduce LongBEL, a document-level normalization framework that leverages full-document

context and previous linking decisions to produce more coherent predictions across biomedical documents.

Together, these contributions move from application to methodology. The first contribution evaluates the practical value of clinical information extraction in a real CDW setting, while the next two focus on improving BEL, by increasing concept coverage and improving document-level consistency.

4 Outline

The manuscript is organized as follows:

1. Chapter 1 introduces the necessary background on biomedical NLP, electronic health record data, and clinical data warehouses.
2. Chapter 2 presents the design and evaluation of a complete end-to-end pipeline for extracting structured information from discharge summaries.
3. Chapter 3 addresses the limited coverage of biomedical entity linking methods by introducing SynCABEL.
4. Chapter 4 focuses on improving document-level consistency in biomedical entity linking through the LongBEL approach.

5 Associated Publications and Preprints

This thesis is based on three research works, corresponding to Chapters 2, 3, and 4. The work presented in Chapter 2 was published in *JMIR Medical Informatics* in 2025 (Remaki et al., 2025). The work presented in Chapter 3 has been presented at IJCAI–ECAI 2026 and published (Remaki et al., 2026a). The work presented in Chapter 4 is currently under review for EACL 2027 and is available as an arXiv preprint (Remaki et al., 2026b).

CHAPTER 1

Background

This chapter provides the background needed to understand the rest of the thesis. It is intended both for readers familiar with data science but less familiar with healthcare data, and for readers familiar with clinical practice but less familiar with natural language processing (NLP). The first part explains what real-world hospital data are, how they are organized in clinical data warehouses, and why they can be difficult to reuse for research. The second part introduces the main NLP tasks used to transform clinical notes into structured information, including named entity recognition, entity qualification, and biomedical entity linking.

1 Real-World Data in Healthcare

Routine-care data become research-ready only after several transformations.

Healthcare systems produce large amounts of digital data during routine care, but these data do not become research evidence simply because they exist. They must be documented, extracted, transformed, standardized, and interpreted within a technical and organizational context. This section follows this trajectory: it first presents electronic health record (EHR) data, then explains how such data are consolidated into clinical data warehouses (CDWs) and standardized for research through common data models such as the Observational Medical Outcomes Partnership (OMOP) (Hripcsak et al., 2015), and finally discusses the biases that can arise along this path.

1.1 EHRs as Real-World Data

EHRs provide a partial view of patient care.

EHR data are commonly reused for observational research because they provide rich real-world information captured during routine care. However, they should not be interpreted as a complete and direct representation of a patient's health state (Callahan et al., 2020). They primarily document interactions with the healthcare system: a clinical

fact appears in the EHR only if the patient receives care, if the information is observed, measured, or reported, and if it is then documented, coded, and stored in an accessible system. EHR data therefore reflect both patient physiology and healthcare processes. For example, whether a laboratory test is ordered, when it is ordered, and how often it is repeated may itself carry information about clinical decisions, workflows, and patient severity (Agniel et al., 2018).

They are powerful for research, but methodologically difficult.

The main strength of EHR-derived real-world data is that they are collected at large scale, often over long periods of time, and in settings close to routine clinical practice. This makes them useful for studying large patient populations, repeated measurements, multimodal clinical information, and groups of patients who may be less represented in clinical trials (Callahan et al., 2020). However, their reuse is methodologically difficult because they were not primarily produced for research. As illustrated in Figure 1.1, EHR data are produced through multiple care activities, such as consultations, laboratory testing, imaging, pharmacy workflows, and surgery, before being consolidated and made available in a clinical data warehouse. Each step may affect what is recorded, how it is represented, and whether it can later be reused for research. EHR-based studies may therefore be affected by missingness, irregular sampling, temporal complexity, changes over time, and variability across sites (Al-Sahab et al., 2024). Particular care is needed when defining cohorts, variables, and time windows.

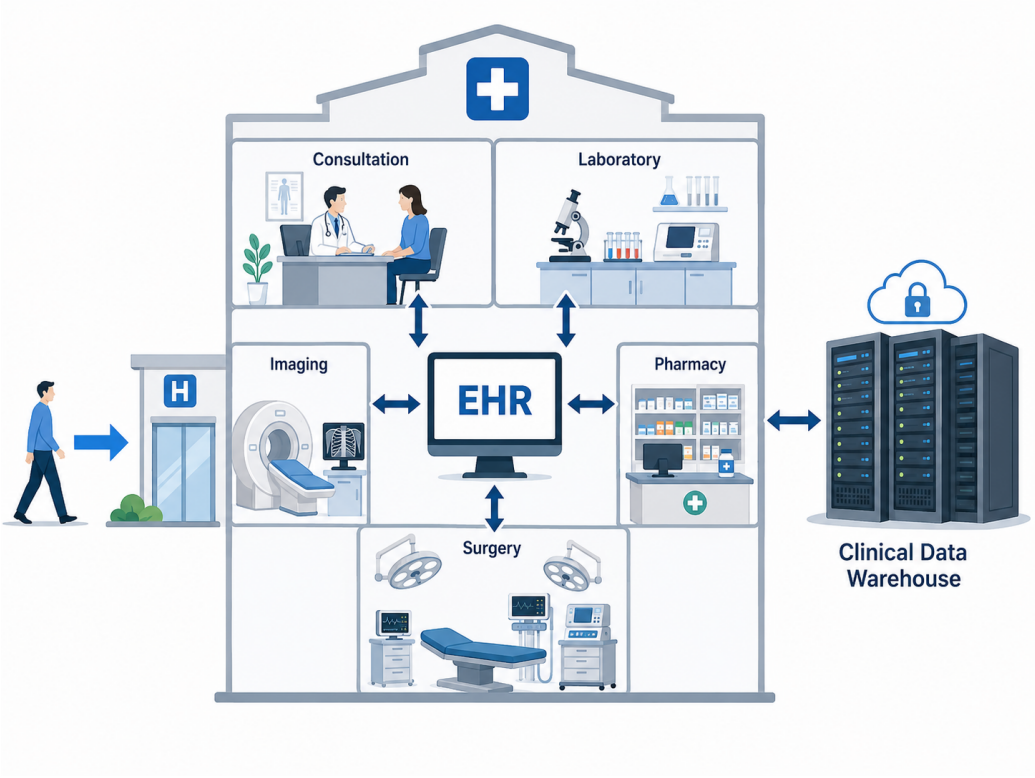


Figure 1.1: Simplified overview of an EHR-based hospital information system. A non-exhaustive set of care activities produces data recorded in the EHR, which can then be transferred to a clinical data warehouse (CDW) for secondary use. Inspired by Laforet (2017).

1.2 From Care Delivery to an OMOP Research Database

Routine care generates heterogeneous clinical data.

Before becoming research data, EHR-derived data are produced in the course of care delivery: patients are admitted, examined, prescribed treatments, monitored, coded, and discharged. In some hospital networks, these activities are distributed across many separate information systems; in others, such as Greater Paris University Hospitals (Assistance Publique–Hôpitaux de Paris, AP-HP), much of the clinical documentation is organized around a shared EHR environment. However, this does not make the resulting data homogeneous. Within the same operational environment, data are generated through different clinical workflows, forms, coding practices, terminologies, and levels of structure. CDWs are therefore needed to consolidate these heterogeneous clinical information and make them available for secondary use (Doutreligne et al., 2023). The AP-HP CDW data catalogue illustrates this diversity in practice, with data spanning administrative information, clinical events, prescriptions, laboratory results, and clinical reports (Daniel et al., 2020). Thus, a research database is not the direct continuation of care delivery, but the result of a process that gathers, reorganizes, and transforms EHR data for secondary use.

A CDW is a transformation infrastructure, not only a storage system.

The construction of a CDW involves several steps between source routine-care data and the final research platform. Source data must first be copied from operational systems. They are then transformed through multiple technical and organizational processes, including procedures for harmonization, quality control, and privacy protection. The transformed data must also be stored in a form that supports reuse, and finally made accessible through tools or platforms that allow authorized users to define cohorts, extract variables, and analyze data (Doutreligne et al., 2023). Studies based on AP-HP data show that these data flows, from hospital information systems to the CDW and then to the analysis platform, can affect data completeness and therefore the final study population (Priou et al., 2024). Thus, the research dataset is already the result of methodological choices made during data extraction, transformation, storage, and access.

This architecture is not specific to one institution. Figure 1.2 shows the Stanford STARR data science ecosystem as an example of how heterogeneous real world data (RWD) are transformed into research datasets and tools. Although STARR and the AP-HP CDW differ in institutional context, the figure illustrates a general CDW logic: routine-care data are ingested, processed, linked, standardized, de-identified, and made available for research use.

Standardization is both a research layer and a transformation step.

Even after integration into a CDW, hospital data may remain organized according to local schemas, formats, and vocabularies. A widely used example of standardization for observational health research is the OMOP common data model, developed within the Observational Health Data Sciences and Informatics (OHDSI) community. OMOP adds a common research layer by standardising observational health data and supporting large-scale analyses across heterogeneous databases (Hripcsak et al., 2015). This standardization also relies on OHDSI Standardized Vocabularies, which provide a com-

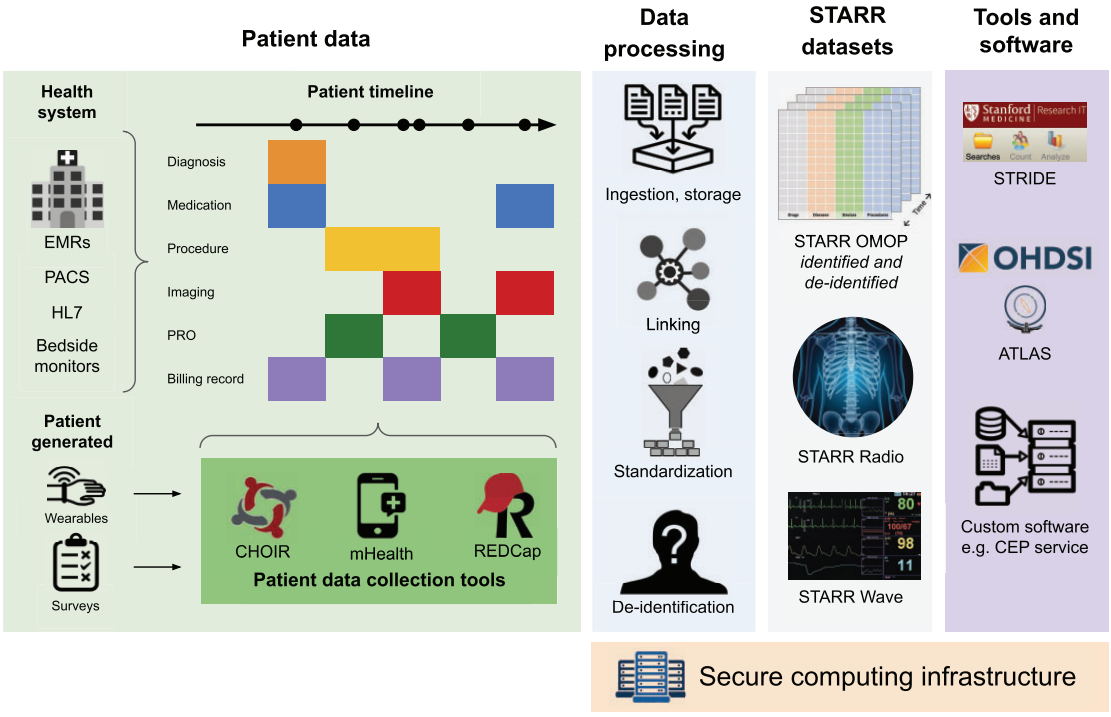


Figure 1.2: Example of a clinical data warehouse pipeline. The Stanford STARR ecosystem illustrates the transformation of heterogeneous patient data sources into research datasets, analytical tools, and software within a secure data science infrastructure. Although this example comes from Stanford, the same conceptual steps are relevant to other large hospital CDWs, including the AP-HP CDW. Reproduced from Callahan et al. (2023).

mon reference system for representing clinical concepts across sites (Reich et al., 2024). Producing an OMOP database therefore requires harmonization rather than a simple copy: source data must be profiled, structurally mapped to OMOP tables, semantically mapped to standard concepts and checked for quality (Henke et al., 2024). Empirical work on OMOP transformation shows that much of the source information can be preserved, but also that mapping difficulties and residual information loss remain domain-dependent, especially for drug prescriptions and laboratory test measurements (Papez et al., 2021). OMOP therefore makes EHR-derived data more reusable, but the resulting database should still be understood as the output of explicit modeling, transformation, and vocabulary-mapping choices. OMOP is a representative example of standardization, but other common data models, terminologies, and interoperability standards exist, and the dominant framework may evolve over time.

The AP-HP CDW also illustrates the need for analysis-oriented transformation tools.

The AP-HP CDW provides a concrete example of the path from routinely collected hospital data to research-usable databases, including OMOP-standardized data. The AP-HP CDW has been developed since 2015 and was made available at scale in 2017 within a secured infrastructure for secondary use (Daniel et al., 2022). Earlier work describes the aggregation of EHR data from AP-HP hospitals, the deployment of a common

EHR environment, and the use of data marts, cohort builders, and notebooks to support research and operational applications (Daniel et al., 2019). However, making data available in a CDW is not sufficient to make analyses easy, reproducible, or clinically meaningful. Beyond the database itself, the AP-HP data science team has developed reusable software tools to help researchers implement, validate, and share algorithms. For example, *eds-scikit* (Petit-Jean et al., 2022) provides a framework for OMOP-based phenotyping and analysis, *bioclean* supports the harmonization of laboratory measurements, and *edsnlp* (Wajsburt et al., 2022) provides reusable components for developing and sharing NLP algorithms on clinical notes. These tools illustrate that secondary use does not stop at data storage or OMOP conversion: it also depends on a software layer that makes algorithms more reproducible, comparable, and reusable across projects. The next subsection discusses why these successive transformation steps can also introduce biases.

1.3 Biases in CDW Data

Biases arise at different stages.

The previous subsection followed the route from source routine-care data to research-ready platform. Along this route, biases can originate from the source data themselves or from the extraction and transformation pipeline. In this subsection, we focus on four important mechanisms: medico-economic coding practices, temporal changes in data availability, patient deduplication, and missingness introduced by source-data and Extract, Transform, and Load (ETL) processes.

Medico-economic coding is useful, but not neutral.

A major part of structured hospital data comes from diagnostic and procedure codes, such as the International Classification of Diseases, Tenth Revision (ICD-10) for diagnoses (World Health Organization, 2004) and, in France, the *Classification commune des actes médicaux* (CCAM) for procedures (Trombert-Paviot et al., 2003). For example, the ICD-10 code E11 indicates type 2 diabetes, while the CCAM code ZBQK002 indicates a chest X-ray. These codes are essential for describing hospital activity and are widely reused for cohort selection, phenotyping, and outcome definition. However, they are not produced for research. In the French context, the national medico-administrative hospital coding system (*Programme de médicalisation des systèmes d'information*) is linked to hospital activity reporting and reimbursement. This means that coding practices may reflect clinical relevance, coding rules, local habits, and financial incentives at the same time (Milcent, 2024). As a result, some conditions may be well captured because they are important for hospital activity or billing, whereas others may be undercoded because they are secondary, ambiguous, or not useful for administrative purposes. A diagnosis code should therefore be treated as a trace of administrative work.

The meaning and completeness of CDW variables can change over time.

Temporal bias is particularly important in multi-hospital CDWs because the data source itself evolves. The deployment of a shared EHR, the activation of new modules, changes in clinical workflows, and improvements in data extraction can all modify the amount and type of data available at a given date.

In the AP-HP CDW, progressive digitization has been shown to create time-dependent

missingness. As a result, naive analyses may detect apparent temporal changes in quality-of-care or epidemiological indicators that partly reflect changes in data availability, rather than true changes in care delivery or disease incidence (Remaki et al., 2023a). In the AP-HP CDW, it has been shown that gradual digitization leads to time-dependent data gaps. As a result, unadjusted analyses may reveal apparent temporal changes in quality-of-care or epidemiological indicators that partly reflect changes in data availability, rather than actual changes in care delivery or disease incidence (Remaki et al., 2023a). The `EDS-TeVa` tool was developed to model this progressive adoption of EHR data over time and across sites, and to support analyses restricted to sufficiently digitized sources (Remaki et al., 2023b). Figure 1.3, produced from AP-HP CDW data, illustrates this problem by showing that several EHR functionalities reached high data availability at different periods, rather than being uniformly available from the beginning of the study window. This type of correction can reduce temporal bias, but it may also reduce the number of eligible patients or observations, and therefore decrease statistical power.

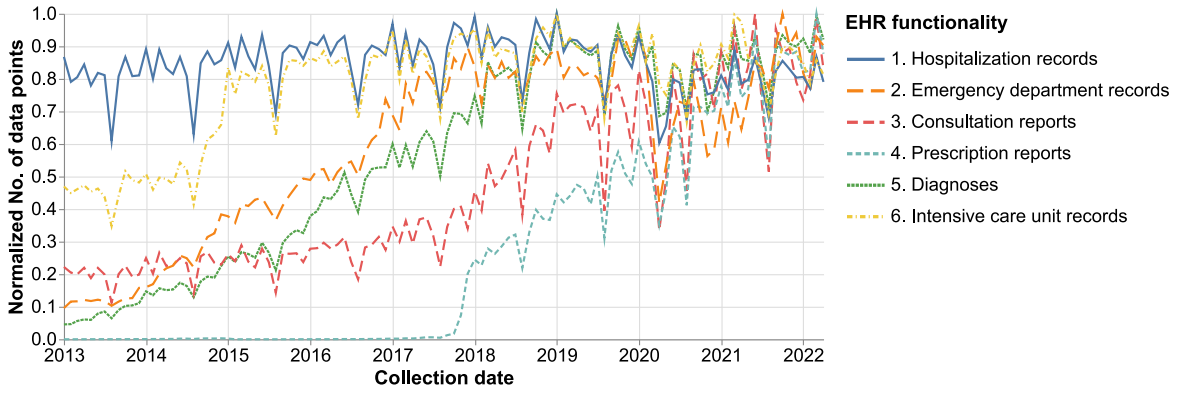


Figure 1.3: Progressive EHR digitization over time in the AP-HP clinical data warehouse. The figure shows the normalized number of data points available for several EHR functionalities between 2013 and 2022, using data from the AP-HP clinical data warehouse. Different functionalities become available at different periods, illustrating why calendar time in a CDW is also a data-availability variable. Reproduced from the `EDS-TeVa` work (Remaki et al., 2023b).

Patient deduplication can affect reconstructed trajectories.

In a large hospital network, the same individual may appear under several local identifiers, and different records must be linked to reconstruct a patient-level trajectory. This deduplication step is necessary, but it can introduce errors. A false split may fragment a patient's history across several identifiers, whereas a false merge may combine data from different individuals. These errors can affect cohort counts, dates of first diagnosis, follow-up duration, exposure histories, and outcomes. They may also be unevenly distributed across patient groups if linkage variables are incomplete or unreliable. For example, record linkage can be affected by biases in dates of birth, including overrepresentation of default dates such as January 1st or December 31st (Fellegi et al., 1969; Harron et al., 2014).

Missingness can come from both source data and ETL processes.

Some information is already missing in the source data because an event was not observed, a test was not ordered, a condition was not coded, or care occurred outside the hospital. This source-level missingness reflects clinical decisions, patient trajectories, and healthcare-system processes (Agniel et al., 2018). Additional incompleteness may then be introduced during ETL, when data are extracted, filtered, transformed, or made available in a research extract. In AP-HP, work on breast cancer care pathways showed how data can be lost between hospital information systems, the CDW, and the analysis platform, reducing the number of patients for whom all required variables are available (Priou et al., 2023). As a result, analyses based only on complete structured variables may select patients not only according to their clinical characteristics, but also according to how completely their data were documented, transferred, and transformed.

Unstructured clinical text can reduce some missingness, but requires NLP.

Structured CDW data are therefore powerful but partial: they provide large-scale, directly queryable variables, while remaining sensitive to coding practices, temporal changes, linkage decisions, and missingness. Clinical narratives can complement these structured fields because they often contain information that was not coded, not standardized, or not available in predefined tables. However, this does not remove all bias. Clinical notes are also affected by temporal changes in documentation, by clinician writing practices, and by the same care processes that shape the rest of the EHR. Moreover, free text is not directly computable: it must first be transformed into structured variables using NLP methods. This transformation is never perfect and must be evaluated carefully, because NLP systems may miss relevant information, extract incorrect mentions, or assign wrong attributes or codes. The next section introduces these NLP methods and the specific difficulties involved in extracting research-ready information from clinical text.

2 Natural Language Processing for Clinical Text

Clinical text contains detailed information, but it must be processed before it can be used as research data.

Clinical narratives record symptoms, diagnoses, treatments, test interpretations, medical reasoning, and contextual information that may be absent from structured fields (Wajsbürt, 2021). However, because they are written in natural language, they cannot be queried or analyzed at scale in the same way as tabular data. The objective of clinical NLP is therefore to transform these texts into structured and clinically meaningful variables. This section first gives a brief history of NLP methods, then focuses on the biomedical information extraction tasks that are central to this thesis: named entity recognition, qualification, relation extraction, and normalization.

2.1 A Brief History of Natural Language Processing

Early NLP systems relied on explicit rules and linguistic categories.

Before the rise of modern neural models, many NLP systems relied on explicit rules to segment, categorize, and interpret text. Part-of-speech tagging illustrates this approach: each word is assigned a grammatical category, such as noun, verb, adjective, or preposition. Transformation-based tagging is a classic rule-based method, where predictions are improved through interpretable correction rules learned from data (Brill, 1995). More generally, classical NLP systems used dictionaries, grammars, and pattern-matching rules to analyze sentences or extract predefined types of information. These methods were relatively interpretable, because their decisions could often be traced back to explicit rules. However, they were also difficult to adapt: a rule that works well in one domain, language, or writing style may fail in another. Although NLP has evolved considerably since then, rule-based methods remain useful in practical applications. Modern libraries such as spaCy still provide matchers and entity rulers, which are especially useful when precise domain-specific patterns must be detected and combined with statistical or neural models (Honnibal et al., 2018).

Neural NLP introduced distributed representations of words.

A major change came with word embeddings, which represent words as vectors learned from large text corpora. Instead of treating words as isolated symbols, these methods place words in a continuous space where words used in similar contexts tend to have similar representations. Models such as word2vec and GloVe made this idea influential by showing that useful semantic and syntactic regularities could be learned from raw text (Mikolov et al., 2013; Pennington et al., 2014). This changed the way text was represented in NLP: words were no longer described mainly through manually designed features, but through numerical representations learned from data. These embeddings were then used as input representations for neural models across many NLP tasks, contributing to a broader shift from feature engineering toward representation learning.

Transformers made word representations contextual and adaptable.

Static word embeddings assign one representation to each word, even though word meaning depends on context. For example, the meaning of *cold* differs between *cold weather* and *common cold*. Transformers addressed this limitation through self-attention, a mechanism that allows the representation of each token, that is, each word or subword unit processed by the model, to depend on the surrounding text (Vaswani et al., 2017). BERT, a transformer-based language model, then showed that large bidirectional models could be pretrained on large corpora and adapted to many downstream NLP tasks (Devlin et al., 2019). This contextual pretraining paradigm was then extended in different directions depending on the available data. Some models were adapted to a specific language, such as CamemBERT for French (Martin et al., 2020). Others were adapted to specific domains: BioBERT was pretrained on biomedical literature, while ClinicalBERT adapted BERT to clinical notes (Lee et al., 2020; Huang et al., 2019). This is particularly important in clinical NLP because hospital narratives contain abbreviations, specialized terminology, fragmented syntax, and documentation conventions that differ substantially from general-language text. In the context of this thesis, an es-

pecially relevant example is EDS-CamemBERT, which adapts CamemBERT to French clinical notes from the AP-HP clinical data warehouse (Dura et al., 2022).

Large language models scaled pretraining and broadened NLP applications.

Large language models (LLMs) pushed the pretraining paradigm further by increasing model size, training data, and the computing resource requirements. Earlier neural NLP systems were often trained or fine-tuned for a specific task, such as classification, sequence labeling, or question answering. In contrast, LLMs provide a more general interface: many tasks can be formulated as natural-language instructions, sometimes with only a few examples. GPT-3 was an important milestone in this shift, showing that a single large autoregressive language model could perform a wide range of NLP tasks in a few-shot setting (Brown et al., 2020). Since then, LLM applications have expanded rapidly, including text generation, question answering, summarization, information extraction, coding, and decision-support systems. This expansion has also reached biomedical NLP, where LLMs are now evaluated on tasks such as biomedical named entity recognition, relation extraction, document classification, question answering, summarization, and text simplification (Chen et al., 2025b). However, in scientific and clinical contexts, the usefulness of these models does not only depend on their flexibility. Their outputs must also be reliable, reproducible, and evaluated for the specific task and setting in which they are used.

In CDWs, NLP is most useful to produce structured and verifiable information.

The history of NLP therefore shows a gradual movement from manually written rules to learned representations, contextual pretraining, and general-purpose language models. Yet, for the secondary use of clinical notes, the practical objective remains specific: free text must be transformed into variables that can be counted, queried, linked with structured EHR data, and reused in clinical studies. This is the role of information extraction. In biomedical and clinical text, information extraction typically involves detecting clinically relevant mentions, qualifying their context, linking them to standardized concepts, and identifying relations between them (Wang et al., 2018; Bazoge et al., 2023; Wajsbürt, 2021). For example, detecting the word *diabetes* is not sufficient: an NLP system must also determine whether diabetes is affirmed or negated, current or past, experienced by the patient or mentioned in family history, and whether it can be mapped to a standardized medical concept. The rest of this section therefore focuses on the core information extraction tasks needed to transform clinical narratives into research-ready data.

2.2 Biomedical Named Entity Recognition

Biomedical named entity recognition identifies text spans that correspond to biomedical concepts.

Biomedical named entity recognition (BioNER) is the task of detecting mentions in text and assigning them to predefined semantic categories (Park et al., 2024). It therefore combines two decisions: identifying the boundaries of the entity mention and determining its semantic type. Depending on the corpus and the application, these entities may include diseases, genes, proteins, chemicals, drugs, species, anatomical entities,

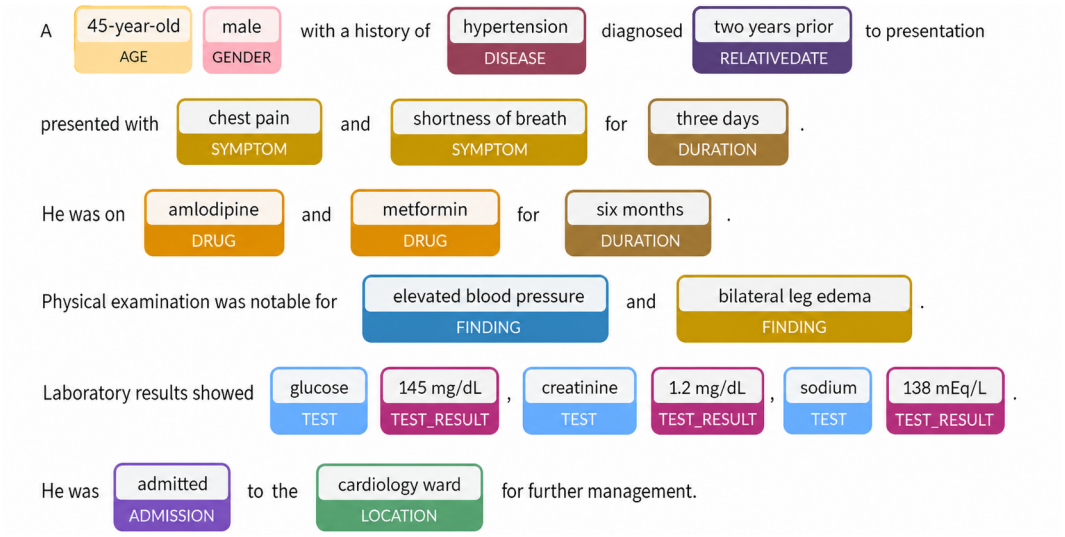


Figure 1.4: Example of biomedical named entity recognition on a synthetic clinical text. Relevant spans are identified and assigned semantic categories, such as diseases, symptoms, drugs, laboratory tests, or demographic information. The example illustrates that named entity recognition requires both selecting the correct span boundaries and assigning the correct entity type.

symptoms, laboratory tests, measurements, clinical events, or any other category considered relevant for the extraction task. Figure 1.4 illustrates this principle: NER is not simply keyword search, but the structured annotation of the right span with the right category.

BioNER can be formulated either as token labeling or span prediction.

As illustrated in Figure 1.5, entity recognition can be formulated either as token labeling or as span prediction. A common formulation represents entity recognition as a token-labeling task: instead of directly predicting complete entities, the system assigns a tag to each token and reconstructs entity spans from the predicted tag sequence. In the classical BIO scheme, tokens are marked as beginning, inside, or outside an entity (Ramshaw et al., 1995). More detailed variants such as BIOUL additionally distinguish the last token of a multi-token entity and entities composed of a single token (Ratinov et al., 2009). This formulation is simple, efficient, and naturally fits models that predict one label per token. In multi-label settings, separate BIOUL tag sequences can also represent overlapping or nested entities when they belong to different categories. However, standard token-labeling schemes remain limited when nested mentions belong to the same category, because the same token would need to carry several boundary roles for the same entity type (Wajsbürt, 2021). Entity recognition can therefore also be formulated at the span level: instead of labeling each token, the model directly proposes candidate text spans and assigns an entity type to each span. Span-based formulations are particularly useful when entities overlap or when several candidate spans must be compared directly, whereas token-labeling formulations remain widely used because they are simple to train and efficient to decode.

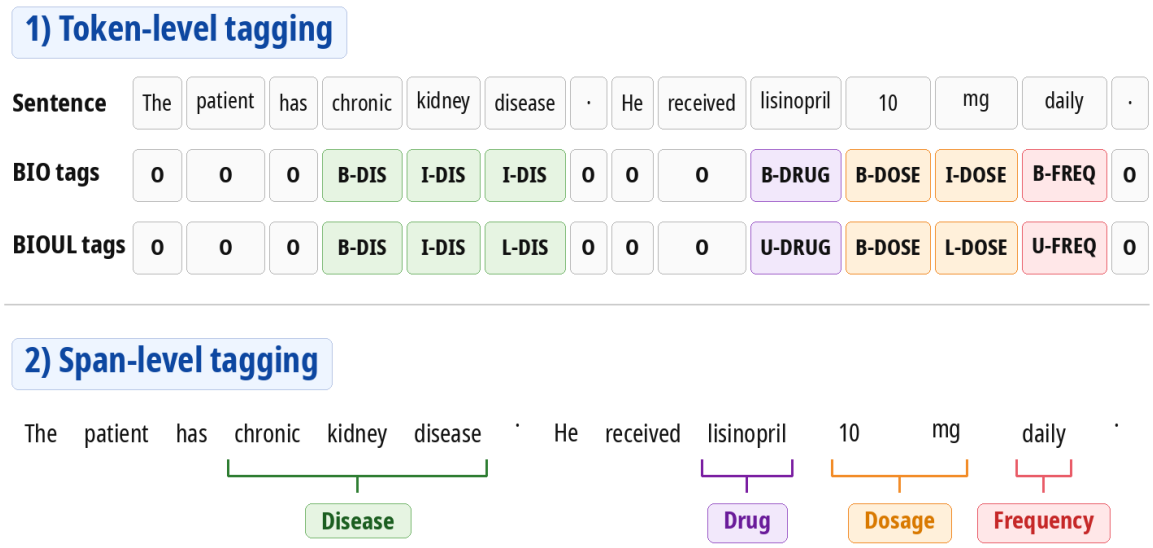


Figure 1.5: Comparison of common formulations for named entity recognition. In BIO tagging, each token receives a label indicating whether it begins an entity, continues an entity, or lies outside any entity. BIOUL adds more explicit boundary information by distinguishing beginning, inside, last, unit-length, and outside tokens. In span prediction, the system directly detects and classifies complete text spans instead of reconstructing entities from token-level labels.

Rule-based BioNER relies on knowledge and manually defined patterns.

Early BioNER systems identified entities by matching text against dictionaries, biomedical terminologies, gazetteers, lexical rules, regular expressions, and domain-specific naming conventions (Campos et al., 2012). The intuition is simple: if the target entities belong to a known vocabulary, such as diseases, species, drugs, genes, or proteins, then part of the task can be addressed by searching for known names, synonyms, variants, and characteristic surface forms. Specific dictionary-based systems were first developed for particular biomedical entity types. ProMiner, for example, was designed to recognize gene and protein names in biomedical text (Hanisch et al., 2005), while LINNAEUS focused on species name recognition (Gerner et al., 2010). Broader terminology-driven systems then aimed to cover a wider range of biomedical and clinical concepts. MetaMap is a classical example of this approach, mapping biomedical text to concepts from the Unified Medical Language System (UMLS) (Aronson et al., 2010). In clinical NLP, cTAKES applies a similar logic by combining linguistic preprocessing with dictionary-based clinical concept extraction (Savova et al., 2010). Rule-based and dictionary-based methods remain widely used because they are transparent, easy to inspect, and fast to process, making them well suited for large-scale processing. However, they often require knowledge resources and manually designed rules, which can be time-consuming to develop. They also do not always generalize well: they can miss unseen variants, local abbreviations, spelling errors, or institution-specific expressions, which often leads to high maintenance costs when the corpus, language, specialty, or entity schema changes (Campos et al., 2012). More recent NLP frameworks have made rule-based approaches easier to implement, combine, and share. General-purpose libraries such as spaCy provide practical components for phrase matching and rule-based entity annotation, while EDS-NLP

provides a framework for developing and sharing rule-based pipelines for French clinical notes (Wajsburt et al., 2022).

Supervised BioNER moved entity recognition from manually defined patterns to learned representations.

After rule-based systems, BioNER was increasingly formulated as a supervised sequence-labeling task, where the model predicts a label for each token and reconstructs entity spans from the predicted label sequence. Conditional random fields (CRFs) were an important step in this direction because they model dependencies between neighboring labels and can produce coherent entity boundaries (Lafferty et al., 2001). Neural models then replaced many manually designed features with learned representations. BiLSTM-CRF architectures, for example, combined word-level representations, character-level information and contextual sequence modeling, making them effective for recognizing variable biomedical mentions (Huang et al., 2015; Lample et al., 2016; Ma et al., 2016). Transformer-based encoders later extended this logic by using pretrained contextual representations that can be fine-tuned for token- or span-level prediction. Biomedical encoders such as BioBERT and PubMedBERT became strong supervised baselines for BioNER (Lee et al., 2020; Gu et al., 2021). In French clinical text, the same principle has been applied through the fine-tuning of CamemBERT-EDS for extracting laboratory tests and drugs from discharge summaries (Remaki et al., 2025). Overall, deep-learning BioNER methods shift the effort from manually specifying entity forms to learning them from annotated examples, with approaches ranging from single-task neural models to transfer-learning and hybrid systems (Song et al., 2021). Compared with rule-based methods, they can better exploit context and variation in wording. They can achieve strong performance while remaining relatively fast to process, making them suitable for large-scale processing. However, they require annotated training data, they are less directly interpretable and the fine-tuning and inference can also require GPU resources.

Large language models and open-vocabulary methods make BioNER more flexible, but not automatically more reliable.

Recent work has explored BioNER as a prompting, few-shot or instruction-following task. Instead of training a model only for one fixed label set, the extraction target can be described through natural-language instructions or prompts. BioNER-LLaMA illustrates this direction by transforming BioNER from a sequence-labeling task into an instruction-guided generation task and evaluating this approach on disease, chemical, and gene recognition datasets (Keloth et al., 2024). A related direction is open-vocabulary NER with lightweight encoder-based models. GLiNER uses natural-language entity type descriptions with a bidirectional encoder, allowing parallel span extraction without the cost of autoregressive generation (Zaratiana et al., 2024). OpenBioNER-v2 follows a similar type-description paradigm, but targets BioNER with lightweight models designed for zero-shot extraction (Cocchieri et al., 2026). These approaches are attractive because they combine part of the flexibility of prompting with models that are smaller and more directly adapted to span detection. However, BioNER remains a strict span-level task: exact boundaries, stable labels, reproducible outputs, and parsable predictions are essential. In clinical notes, recent comparisons suggest that instruction-tuned LLMs can improve performance over BERT-style baselines for NER and relation extraction, especially in limited-data or unseen-data settings, but with substantially

higher computational cost and lower throughput (Hu et al., 2026). Broader biomedical NLP benchmarks also show that zero-shot and few-shot LLMs are not consistently superior to fine-tuned encoder-based models for information extraction tasks and may produce missing, inconsistent, or hallucinated outputs (Naguib et al., 2024; Chen et al., 2025b). These methods should therefore be seen as promising tools for flexible and low-resource extraction, rather than as automatic replacements for validated supervised BioNER pipelines.

Table 1.1: Synthesis of BioNER model families according to their strengths and limitations

Model family	Examples	Strengths	Limitations
Rule-based	Dictionaries MetaMap cTAKES EDS-NLP	+ Interpretable + No training data + Fast processing	– Poor generalization – Need prior knowledge – Maintenance cost
Supervised encoder	BioBERT PubMedBERT CamemBERT-EDS	+ Best performance + Fast and lightweight	– Annotated data – Need GPU
Open-vocabulary encoder	GLiNER OpenBioNER-v2	+ Highly flexible + Fast and lightweight + Zero-shot	– Lower performance – Sensitive to input
Instruction-tuned LLM	BioNER-LLaMA	+ Highly flexible + Zero-/few-shot	– Lower throughput – Sensitive to input – Inconsistent output

BioNER evaluation depends on both entity boundaries and entity types.

BioNER systems are usually evaluated with precision, recall, and F1-score at the entity level. Under strict evaluation, a prediction is counted as correct only when the predicted span boundaries and the predicted entity type match the reference annotation (Tjong Kim Sang et al., 2003). This makes evaluation more demanding than token-level accuracy: predicting only part of an entity, adding an extra word, or assigning the wrong type is counted as an error. Because BioNER combines mention detection and type classification, such errors must be interpreted in relation to the target application: a partial span may be acceptable for document retrieval, but insufficient when the extracted mention must be normalized, counted, or used as a research variable (Nadeau et al., 2007). More importantly, even a correctly detected entity is not yet a complete clinical fact. A mention may be negated, uncertain, historical, hypothetical, or refer to a family member rather than to the patient. BioNER therefore provides the textual units of interest, but clinical information extraction requires additional steps to qualify their context and link them to standardized biomedical concepts.

2.3 Clinical Entity Qualification

Entity qualification turns detected mentions into clinically interpretable facts.

Named Entity Recognition identifies spans of text that correspond to clinical entities, such as diseases, symptoms, procedures, or drugs. However, the presence of a clinical word in a report does not necessarily mean that the corresponding clinical fact is true for the patient. For example, the mention *fever* in "no fever" should not be interpreted as an observed symptom; *pneumonia* in "possible pneumonia" refers to a diagnostic hypothesis; *diabetes* in "history of diabetes" describes a past condition; and *breast cancer* in "mother had breast cancer" concerns a family member rather than the patient. Clinical entity qualification addresses this issue by assigning contextual attributes to detected entities. Similar operations are also described in the clinical NLP literature as *assertion classification*, *attribute detection*, or *modifier prediction*. In this thesis, we use *entity qualification* as an umbrella term for these tasks. Technically, it can be formulated as one or several classification tasks applied to each mention, where the system predicts labels such as negation, certainty, temporality, or family from the mention and its context (Uzuner et al., 2011; Harkema et al., 2009). In this sense, qualification is the step that transforms a textual mention into a clinically usable assertion.

Qualification has often been addressed with interpretable rule-based methods.

A large part of clinical qualification has historically been addressed with rule-based methods. These systems search for lexical cues such as "no", "without", "history of", "possible", "rule out", or "family history of", and then determine which entity mention falls within the scope of the cue. This approach was first popularized for negation detection with NegEx, a regular-expression-based algorithm designed to identify negated findings and diseases in discharge summaries (Chapman et al., 2001). It was then extended by ConText, which used a similar surface-based strategy to detect whether clinical conditions are negated, hypothetical, historical, or experienced by someone other than the patient (Harkema et al., 2009). These rule-based approaches remain attractive in clinical NLP because they are interpretable, easy to adapt to local terminology, and often effective for well-defined contextual phenomena. For instance, in French clinical NLP, EDS-NLP mainly proposes rule-based pipeline components for qualification tasks such as negation, hypothesis, and family-context detection (Wajsburt et al., 2022).

Learning-based methods treat qualification as supervised classification.

Qualification can also be learned from annotated examples. In this setting, each entity mention is associated with a gold contextual label, and the model learns to predict this label from the surrounding text. The Informatics for Integrating Biology and the Bedside (i2b2) and the U.S. Department of Veterans Affairs organized a challenge in 2010 that played an important role in formalizing this view of clinical assertion classification: given medical concepts, systems had to classify their assertion status as present, absent, possible, conditional, hypothetical, or not associated with the patient (Uzuner et al., 2011). Systems developed in this setting often combined machine-learning classifiers with engineered linguistic and contextual features. For example, i2b2 systems used features derived from lexical cues, syntax, local context, section information, and existing clinical NLP tools (Bruijn et al., 2011; Kim et al., 2011). More recent work has evaluated medical language models for the same type of task, for example by classifying clinical mentions as present, absent, or possible (Aken et al., 2021). This evolution does not remove the need for clinical validation: qualification models are still trained

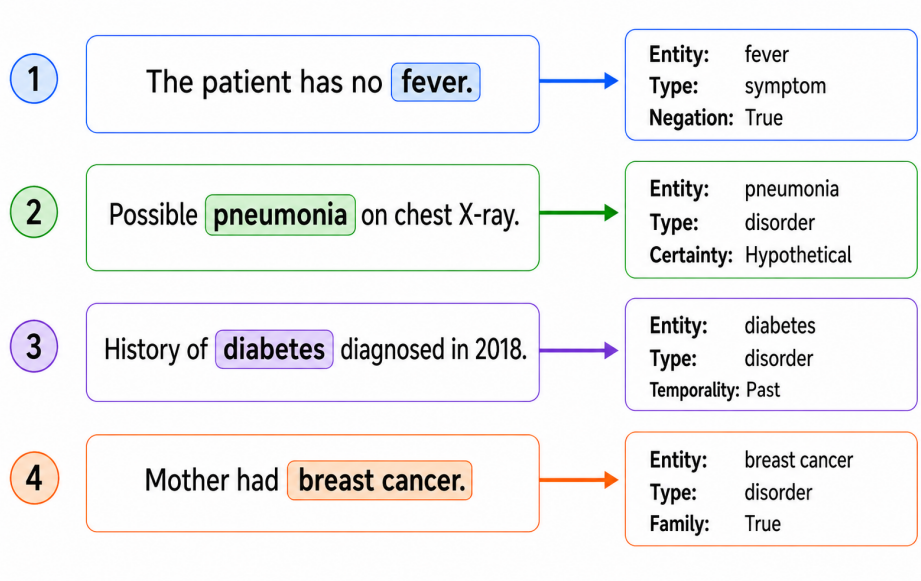


Figure 1.6: Illustration of clinical entity qualification. BioNER detects clinical mentions, while qualification assigns contextual attributes that determine whether the mention is negated, hypothetical, historical, or related to family history.

and evaluated on annotated corpora, and their performance may depend on the type of documents and annotation scheme used.

Table 1.2: Synthesis of clinical entity qualification model families according to their strengths and limitations

Model family	Examples	Strengths	Limitations
Rule-based	NegEx ConText EDS-NLP	+ Interpretable + No training data + Fast processing	– Poor generalization – Needs manual rules – Maintenance cost
Supervised encoder	BioBERT CamemBERT-EDS	+ Uses context + Best performance + Adaptable	– Annotated data – Less interpretable

Evaluating qualification requires NLP metrics and downstream validation.

At the NLP level, qualification is usually evaluated by comparing the predicted attribute of each entity mention with manual annotations, using standard metrics such as precision, recall, and F1-score (Uzuner et al., 2011). These scores can be computed separately for each qualifier, for example negation, uncertainty, temporality, or family context. However, the importance of qualification is best understood through its downstream impact. A negated disease should not be counted as a diagnosis, a disease mentioned in a family-history context should not be attributed to the patient, and a past condition may not define the same phenotype as an active one. This is particularly important in CDWs, where studies have shown that detecting negation and family-history

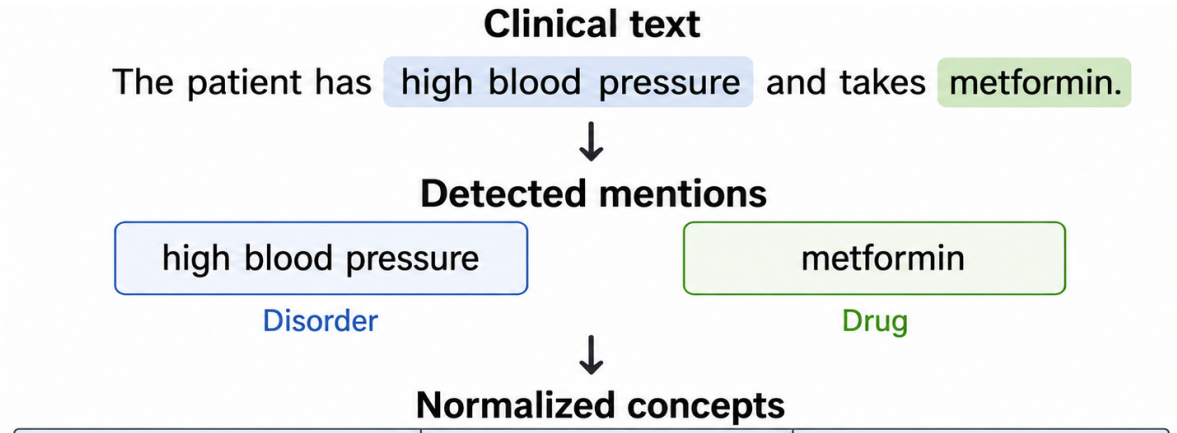
context can substantially affect the identification of patient cohorts from narrative reports (Garcelon et al., 2017). Qualification therefore helps transform extracted mentions into more reliable clinical variables. Yet, even after qualification, the entity is still expressed through the wording used by the clinician. The same concept may appear under different surface forms, abbreviations, or spelling variants, and these variants must be mapped to a shared identifier to make NLP outputs comparable across documents and databases. This mapping step is the objective of entity normalization, which is introduced in the next subsection.

2.4 Biomedical Entity Linking (Entity Normalization)

Biomedical entity linking maps textual mentions to standardized concepts.

BioNER identifies spans of text that correspond to clinically relevant entities, such as disorders, drugs, procedures, or anatomical locations. However, a detected span is still only a surface form. The same medical concept can be expressed in many different ways: for example, *high blood pressure*, *hypertension*, and *hypertensive disease* may all refer to the same underlying disorder. Conversely, the same surface form may be ambiguous: an abbreviation can refer to different concepts depending on the clinical context. Biomedical entity linking (BEL) addresses this problem by mapping textual mentions to standardized concepts in a controlled vocabulary or terminology (Ruas et al., 2026; Chen et al., 2026; French et al., 2023). These resources provide stable identifiers, preferred names, synonyms, and sometimes semantic types or relations between concepts. Important examples include the UMLS, which integrates many biomedical terminologies such as Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), Medical Subject Headings (MeSH), Anatomical Therapeutic Chemical (ATC) and ICD (Bodenreider, 2004; Donnelly, 2006; Lipscomb, 2000; World Health Organization, 2004). Because this mapping step normalizes variable surface forms into shared identifiers, it is also referred to as biomedical entity normalization or biomedical concept normalization. In this thesis, we use *entity linking* and *normalization* interchangeably to refer to this task, and the terms *knowledge base* (KB), *terminology*, and *vocabulary* broadly and interchangeably to refer to resources that provide standardized biomedical concepts and identifiers. Figure 1.7 illustrates this task on a simple clinical sentence. Figure 1.7 illustrates this task on a simple clinical sentence.

This step is especially important in CDWs, where the objective is not only to detect clinical mentions in text, but also to transform them into standardized variables that can be reused for research. BEL provides this normalization layer by mapping textual mentions to standardized identifiers in controlled KBs. A note may mention a broad disorder such as *diabetes*, a more specific disorder such as *type 2 diabetes mellitus*, or an abbreviated form such as *T2DM*. Without BEL, these mentions remain separate textual strings. With BEL, they can be represented by standardized identifiers, or by related concepts at different levels of granularity, and therefore queried, counted, compared across documents, and combined with structured data. This is one reason why BEL is a central step for transforming clinical text into reusable research data (Wajsbürt, 2021). The biomedical domain also has the particularity of relying on many specialized KBs, such as UMLS, SNOMED CT, MeSH, ICD, or ATC, which provide the reference concepts to which textual mentions can be linked (Bodenreider, 2004; Donnelly, 2006; Lipscomb, 2000; World



Mention	Concept code	Concept name
high blood pressure	C0020538	Hypertensive disease
metformin	C0025598	Metformin

Figure 1.7: Example of biomedical entity normalization. After entity mentions have been detected in the clinical text, each mention is mapped to its corresponding standardized concept code in a KB and concept name. This step makes different textual forms computable as the same biomedical concept.

Health Organization, 2024). The task remains technically challenging because biomedical mentions may appear under many surface forms, including synonyms, abbreviations, spelling variants, and locally preferred expressions. These challenges are even stronger in languages other than English, where KB coverage, annotated corpora, and pretrained clinical resources are often more limited. This has been shown for French BEL, where fewer synonyms are available than in English for many UMLS concepts (Névéol et al., 2018).

Technically, BEL consists of choosing the correct concept from a KB.

A simple way to formulate BEL is to consider a mention m, its textual context x, and a target KB $\mathcal{K}$ containing possible concepts. The objective is to select the concept c that best matches the mention:

$$\hat{c} = \arg\max_{c \in \mathcal{K}} s(m, x, c),$$

where $s(m, x, c)$ is a score measuring the compatibility between the mention, its context, and the candidate concept. This closed-set formulation assumes that a suitable concept exists in $\mathcal{K}$, although some mentions may have no corresponding concept in the target KB. In practice, $\mathcal{K}$ can be very large, so many systems do not compare the mention with all concepts using a complex model. Instead, they first retrieve a small set of plausible candidates, and then rank these candidates to select the final concept (Ruas et al., 2026; Chen et al., 2026). This two-step view is useful because it separates two sources of error: the correct concept may be absent from the candidate list, or it may be retrieved but ranked below an incorrect concept. More recent generative methods formulate the same problem differently, by asking a model to generate a concept name or code directly, but

the output still has to correspond to a valid entry in the target KB.

Early BEL systems matched mentions to synonyms in biomedical KBs.

Early BEL systems were mainly based on biomedical KBs. Their core idea was simple: compare the mention found in the text with the concept names and synonyms stored in the KB, then select the best matching concept. To make this comparison more robust, these systems used lexical rules such as spelling normalization, abbreviation expansion, approximate string matching, and semantic-type filtering. MetaMap is a classical example of this approach: it maps biomedical text to UMLS concepts by generating lexical variants, retrieving candidate concepts, and scoring them with linguistic and semantic evidence (Aronson et al., 2010). In clinical NLP, cTAKES follows a similar KB-based logic as part of a broader pipeline for detecting and normalizing clinical concepts (Savova et al., 2010). Later tools such as QuickUMLS and scispaCy UMLS linker made biomedical concept matching easier to apply at scale, using efficient approximate matching and ready-to-use UMLS linking components (Soldaini et al., 2016; Neumann et al., 2019). These KB-based systems remain important because they are transparent, require little or no task-specific annotated data, and are directly grounded in expert-curated biomedical resources. They are also useful as candidate generators in more complex BEL pipelines. However, their performance depends strongly on the names and synonyms available in the KB. They can fail when a clinical mention is abbreviated, misspelled, locally expressed, or very different from the corresponding concept name. They can also struggle when several concepts have similar names and the correct choice depends on context. These limitations motivated later methods that learn mention–concept similarity instead of relying only on handcrafted lexical rules.

Bi-encoder methods learn representations of mentions and concepts.

Neural BEL methods replaced handcrafted matching rules with learned representations. In a bi-encoder formulation, the mention and the candidate concept are encoded separately: one encoder represents the mention, and another represents the concept name or its synonyms. The model then links the mention to the concept whose representation is closest in the embedding space. This is efficient because concept representations can be computed once for the whole KB, and new mentions can then be linked by nearest-neighbor search. Several influential biomedical BEL methods follow this idea. BioSyn combines sparse lexical similarity with dense neural representations and uses synonym marginalization, so that all names of the same concept can contribute to the learning objective (Sung et al., 2020). SapBERT makes this idea even more direct by using UMLS synonymy during pretraining: names that refer to the same concept are learned to be close in the embedding space (Liu et al., 2021a). CODER extends this representation-learning approach by using relations between biomedical concepts in the KB (Yuan et al., 2022c). BERGAMOT follows the same general direction graph-augmented objectives from the UMLS KB to improve biomedical entity representations (Sakhovskiy et al., 2024). These methods are scalable and well suited to candidate generation, but they compare the mention with concept names. They may therefore struggle when the correct concept depends strongly on the surrounding clinical context.

Contextual bi-encoders learn to link mentions from their surrounding text.

Name-based bi-encoders mainly learn from the KB itself: they compare a mention with concept names, synonyms, or definitions. Contextual bi-encoders add another source of information: the sentence or document in which the mention appears. The mention is encoded together with its context, while the candidate concept is encoded separately, using the information available in the KB, such as its name, synonyms, or textual definition. The model then compares both representations in the same embedding space. This keeps the method scalable, because concept representations can be precomputed, while allowing the mention representation to depend on its local context. This idea was popularized in general entity linking by BLINK. A mention encoder represents the sentence containing the mention, while an entity encoder represents each candidate entity from its KB-side description. The mention is then linked to the entity whose vector is closest in the shared embedding space (Wu et al., 2020). Biomedical BEL systems adapt this idea by encoding biomedical mentions in context and candidate concepts from their preferred names, synonyms, or definitions when available (Bhowmik et al., 2021). ArboEL further improves this family of methods by learning mention and entity representations with an arborescence objective that also models mention–mention affinities across documents (Agarwal et al., 2022). KRISSBERT addresses the annotation bottleneck differently: instead of relying only on manually annotated mention–concept pairs, it generates self-supervised contextual mentions from unlabeled biomedical text and a domain KB, then links a test mention to the most similar entity prototype (Zhang et al., 2022). These methods show why context is useful for BEL, especially for ambiguous mentions, but they also highlight a key limitation: learning reliable contextual representations usually requires contextual mention–concept supervision, which is expensive to obtain in biomedical and clinical domains.

Rerankers compare each mention–concept pair more carefully.

After candidate generation, some BEL systems use a reranker to choose the best concept among a small list of candidates. The most common rerankers are cross-encoders: instead of encoding the mention and the concept separately, they encode them together, for example by concatenating the mention context with the candidate concept name, synonyms, or definition. This gives the model a more detailed interaction between the text and the candidate concept, which can help when several candidates are lexically similar or when the mention is ambiguous (Wu et al., 2020; Bhowmik et al., 2021). However, cross-encoders are computationally expensive because they must be applied separately to each mention–candidate pair. They are therefore usually used only after a faster retrieval step.

Generative methods formulate BEL as constrained generation.

Another family of methods treats BEL as a text generation problem. Instead of retrieving and ranking concept vectors, the model receives the mention and its context, then generates a concept name or identifier. This idea was introduced in general entity linking by GENRE, where an autoregressive model retrieves entities by generating their names with constrained decoding (Cao et al., 2021a). This is one of the strengths of the approach: the model does not need to store a dense vector for every candidate entity, and the decoding process is restricted to the KB entries. It was later extended to multilingual entity linking with mGENRE (Cao et al., 2021b). In biomedical BEL,

GenBioEL adapted this formulation by using KB-guided pretraining, synonym-aware fine-tuning, and constrained decoding over concept names from the KB (Yuan et al., 2022b). More recently, ANGEL improved generative BioEL by explicitly training the model with negative samples, so that it learns to distinguish the correct concept from similar but incorrect concepts (Kim et al., 2025). Generative methods are attractive because they provide a direct sequence-to-sequence formulation of BEL and reduce the need for large dense indexes. However, they can still require costly mention–concept supervision or carefully designed KB-based training data, and autoregressive decoding can make inference slower than a simple nearest-neighbor search.

Table 1.3: Synthesis of BEL model families according to their strengths and limitations

Model family	Examples	Strengths	Limitations
Rule-based	MetaMap cTAKES QuickUMLS	+ Interpretable + No training data + Fast retrieval	– Poor generalization – Context-insensitive
Self-supervised bi-encoder	SapBERT CODER BERGAMOT	+ Better generalization + Zero shot + Scalable retrieval	– Context-insensitive – Needs KB embeddings
Contextual bi-encoder	BLINK ArboEL KRISSBERT	+ Context-aware + Strong performance + Scalable retrieval	– Needs finetuning – Needs KB embeddings
Fine-tuned generative model	GENRE GenBioEL ANGEL	+ Context-aware + Best performance	– Need finetuning – Lower throughput

Large language models are used to support several steps of BEL pipelines.

Recent work has explored large language models (LLMs) as flexible components for BEL. One use is reranker: after a first system retrieves possible concepts, an LLM can be prompted to choose the best concept from this candidate list. PromptLink follows this idea with multi-stage prompts for BEL (Xie et al., 2024), and BioELQA formulates BEL as a multiple-choice question answering task, where the model selects the correct concept among candidate answers (Lin et al., 2024). LLMs can also be used before linking. For example, complex mentions can be simplified or rewritten before candidate generation, which helps the linker retrieve better candidates (Borchert et al., 2024). Another use is data augmentation: LLMs can generate additional training examples or improve the quality of training data for BEL (Chen et al., 2025a). This last direction will be developed in Chapter 3, where we introduce SynCABEL as a method for generating contextualized training examples for BEL. Overall, LLMs are not a single replacement for BEL systems, but a useful set of tools for candidate selection, mention rewriting, data augmentation, and annotation support.

Evaluation checks whether the predicted concept matches the reference.

BEL is usually evaluated by comparing the predicted concept code with a reference concept code. In the simplest setting, the mention span is already given, and the system only has to select the correct concept. The main metric is often Recall@1, also called top-1 accuracy, which checks whether the first predicted concept is the gold concept. Recall@k is also useful: it measures whether the correct concept appears among the first k retrieved candidates. This distinction is important in two-stage systems, because an error can occur either because the correct concept was not retrieved, or because it was retrieved but ranked below an incorrect candidate (Bathala et al., 2025). Evaluation is also difficult because biomedical BEL benchmarks are highly diverse. They may differ by language, by document type, by target KB, and by entity type: some datasets focus on drugs, others on genes, disorders, procedures, or broader biomedical concepts. They may also differ in their annotation guidelines. For example, some corpora annotate long composite mentions, whereas others split them into smaller atomic entities; some rely on automatic pre-annotation followed by correction, whereas others are manually annotated from scratch. These choices affect what counts as a correct span, a correct concept, and a comparable system output (Ruas et al., 2026; Chen et al., 2026). Exact concept-code matching is useful because it is clear and reproducible, but it should be complemented by error analysis, especially when normalized mentions are intended to enrich a CDW.

3 Conclusion

CDWs make routine-care data reusable, but not automatically research-ready.

This chapter first introduced the nature of EHR data. They are valuable because they are produced at large scale during routine care, but they are not a direct and complete representation of the patient. They are shaped by clinical workflows, documentation practices, coding rules, information systems, and technical transformation steps. CDWs, and in particular OMOP-standardized databases, make these data easier to query and reuse for research. However, this reuse requires caution: variables may be affected by missingness, temporal changes in data availability, medico-economic coding practices, patient deduplication, and ETL processes. Research-ready data are therefore not simply collected; they are constructed through a sequence of methodological choices.

Clinical text contains complementary information, but requires NLP.

The second part of the chapter focused on clinical natural language processing. Clinical notes contain rich information that is often absent from structured tables, including symptoms, diagnoses, treatments, laboratory interpretations, temporal information, uncertainty, and clinical reasoning. However, free text cannot be directly queried or analyzed at scale. NLP is needed to transform these narratives into structured information. This transformation is not a single task: BioNER detects relevant mentions, qualification determines how these mentions should be interpreted in context, and BEL maps them to standardized concepts. Together, these steps form the basis of an information extraction pipeline able to convert clinical narratives into computable data.

NLP-derived variables must be evaluated in relation to their clinical use.

The chapter also emphasized that NLP outputs are not automatically valid research variables. Each step of the pipeline can introduce errors: a mention can be missed, incorrectly delimited, assigned the wrong qualifiers, or linked to the wrong concept. These errors matter because they can affect cohort selection, phenotyping, and downstream analyses. For this reason, NLP systems must be evaluated not only as technical models, but also as components of a clinical data pipeline. Their usefulness depends on whether the extracted variables are reliable enough for the intended research question and whether they can meaningfully complement structured CDW data.

The next chapter instantiates this pipeline in the AP-HP CDW.

This general background motivates the first applied contribution of the thesis. The next chapter introduces an end-to-end pipeline designed to extract drugs and laboratory tests from discharge summaries and use them jointly with structured data in the AP-HP CDW. This chapter therefore moves from the general principles introduced here to a concrete use case: showing how NLP can help enrich a large hospital CDW with information extracted from real clinical narratives.

CHAPTER 2

Enriching Phenotyping Through Discharge Summary Processing

This chapter presents a first applied contribution of the thesis: an end-to-end natural language processing (NLP) pipeline for extracting laboratory tests and drug treatments from French discharge summaries and combining them with structured electronic health record (EHR) data. The approach is evaluated on patients with immune-mediated inflammatory diseases in a clinical data warehouse (CDW).

1 Introduction

1.1 Context

The previous chapter introduced the general challenge of transforming clinical narratives into structured information that can be reused for research. This chapter now studies this problem in a concrete setting: patient phenotyping in the Greater Paris University Hospitals (Assistance Publique–Hôpitaux de Paris, AP-HP) CDW. Patient phenotyping consists in characterizing patients through observable clinical features, such as diagnoses, symptoms, laboratory test results, and past or current treatments (Richesson et al., 2013). At the scale of a CDW, such phenotypes can support several secondary uses of EHR data, including predictive risk assessment, preselection of patients for therapeutic trials, pharmacovigilance, and the construction of cohorts of clinically comparable patients (Gombar et al., 2019; Callahan et al., 2021; Frankovich Jennifer et al., 2011).

However, building reliable phenotypes from structured data alone remains difficult. Structured tables provide directly queryable information, such as International Classification of Diseases, Tenth Revision (ICD-10) codes, laboratory results, prescriptions, or drug administrations, but they do not always capture the full clinical trajectory of the patient. Some information may be missing because it was produced outside the hospital, because it predates the period covered by the structured database, or because it was documented only in a clinical report. Discharge summaries are therefore a valuable source of complementary information: they synthesize the patient's history, relevant bi-

ological findings, treatments, and follow-up plans, and may contain information that is absent from structured tables (Zheng et al., 2024). Previous studies have shown that combining structured EHR data with information extracted from clinical text can improve cohort identification, clinical prediction, and patient characterization (Elkin et al., 2021; Seinen et al., 2022; Khurshid et al., 2022; Idnay et al., 2025; Fraile Navarro et al., 2023; Moqurrab et al., 2021).

This issue is particularly relevant for immune-mediated inflammatory diseases where the clinical profile of a patient often depends on biological markers and treatment history. In this chapter, we focus on systemic lupus erythematosus, systemic sclerosis, antiphospholipid syndrome, and Takayasu arteritis. For these diseases, autoantibodies, inflammatory markers, immunosuppressive treatments, anticoagulants, biologics, and vaccinations are clinically important variables. Since these variables may appear both in structured data and in discharge summaries, this chapter evaluates whether an end-to-end NLP pipeline can extract, normalize, and combine them to support more complete patient phenotyping.

1.2 Objectives and Contributions

The objective of this chapter is to provide a proof of concept for end-to-end patient phenotyping from EHR data. More specifically, we evaluate whether information extracted from French discharge summaries can enrich structured data for patients with immune-mediated inflammatory diseases. We focus on laboratory tests and drug treatments because they are central to the characterization of these diseases and may be documented both in structured tables and narrative reports.

The overall workflow is summarized in Figure 2.1. Laboratory tests and drug treatments are extracted from structured and unstructured data, then represented using knowledge bases (KBs) so that they can be analyzed jointly. Laboratory tests are mapped to Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) concepts (Donnelly, 2006) through the Unified Medical Language System (UMLS) (Bodenreider, 2004), while drug treatments are mapped to the Anatomical Therapeutic Chemical (ATC) classification (World Health Organization, 2024). The underlying hypothesis is that discharge summaries contain complementary information that can improve patient characterization while remaining sufficiently reliable and standardized for cohort-level reuse.

This objective is evaluated from two complementary perspectives. First, we assess the intrinsic performance of the NLP pipeline on manually annotated discharge summaries, including extraction, qualification, measurement extraction, and normalization. Second, we assess its contribution to phenotyping by applying it to a large AP-HP cohort and comparing structured data alone with structured data enriched by text-derived variables.

The contributions of this chapter are threefold. First, we present an applied information extraction pipeline for French discharge summaries focused on drug treatments and laboratory tests. Second, we provide a detailed evaluation of the pipeline, both at the component and end-to-end levels. Third, we show how text-derived information can be combined with structured EHR data to improve the characterization of patients with systemic lupus erythematosus, systemic sclerosis, antiphospholipid syndrome, and Takayasu arteritis.

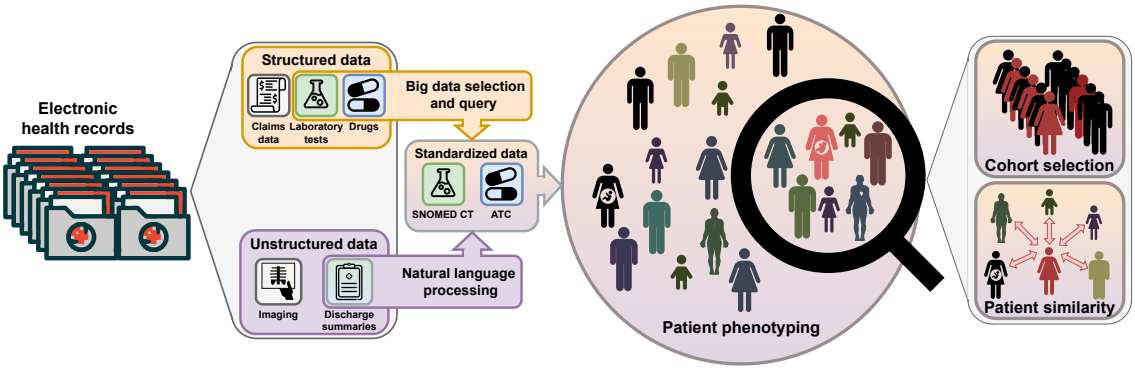


Figure 2.1: Overview of the end-to-end patient phenotyping pipeline. Structured and unstructured data are extracted from electronic health records, enabling large information retrieval, refining cohort selection, and facilitating more robust patient comparisons. ATC: Anatomical Therapeutic Chemical; SNOMED: Systematized Nomenclature of Medicine Clinical Terms.

1.3 Related Work

Chapter 1 introduced the main NLP tasks used to transform clinical text into structured variables, including named entity recognition (NER), qualification, and entity normalization. Here, we focus on previous work using clinical text to improve EHR-based phenotyping, especially through drug and laboratory test extraction.

1.3.1 Clinical NLP Systems for EHR-based Phenotyping

NLP has long been used to complement structured EHR data. Early clinical NLP systems were often complete information extraction pipelines: they detected medical concepts in narrative reports, sometimes assigned qualifiers such as negation or certainty, and mapped mentions to biomedical KBs. Most relied on rule-based or knowledge-based methods, combining lexical matching, regular expressions, and linguistic rules. This is the case for systems such as MedLEE (Friedman et al., 1996), MetaMap (Aronson et al., 2010), HITEx (Zeng et al., 2006), MedEx (Xu et al., 2010), and cTAKES (Savova et al., 2010), although some also included supervised components for specific subtasks.

More recently, clinical and biomedical information extraction has increasingly relied on pretrained language models and, more recently, on methods designed to reduce the need for task-specific supervision. Open-vocabulary models such as GLiNER (Zaratiana et al., 2024) and biomedical variants such as OpenBioNER-v2 (Cocchieri et al., 2026) can identify entities from natural-language descriptions of the target types. In parallel, instruction-tuned large language models have been investigated for biomedical NER in zero-shot and few-shot settings (Keloth et al., 2024; Ge et al., 2026). These approaches offer greater flexibility when adapting to new extraction targets, although recent evaluations show that they do not consistently outperform task-specific supervised models and can incur substantially higher computational costs (Naguib et al., 2024; Chen et al., 2025b; Hu et al., 2026). These methodological developments are discussed in more detail in Chapter 1; in this chapter, representative approaches are used as contemporary reference methods for the NER experiments.

Clinical NLP systems have been used in a wide range of phenotyping and clinical re-

search settings. HITEx was evaluated for extracting principal diagnoses, comorbidities, and smoking status in an asthma study, while MedEx focused on medication names and associated information such as strength, route, and frequency (Zeng et al., 2006; Xu et al., 2010). Other works used text-derived information as features for patient classification. For example, Bejan et al. (2012) trained a supervised model to identify pneumonia from narrative intensive care reports using word n-grams, concept codes, and assertion values. In rheumatoid arthritis, EHR phenotyping studies combined codified data with information extracted from clinical notes to improve case identification (Liao et al., 2010; Carroll et al., 2011). More recent studies have also shown that clinical text can reduce missingness or improve prediction models when combined with structured EHR data (Elkin et al., 2021; Seinen et al., 2022; Khurshid et al., 2022; Idnay et al., 2025).

1.3.2 Drug Extraction from Clinical Text

Drug extraction is one of the most studied tasks in clinical information extraction. Drug mentions are important for phenotyping because they provide information about disease management, disease severity, and patient risk. However, extracting only the drug name is usually not sufficient. A complete drug description also includes information such as strength, dose, route, frequency, duration, or form. These elements describe *what* treatment is mentioned and *how* it is prescribed or administered. Drug extraction systems therefore often go beyond drug name recognition and also extract these associated concept codes (Xu et al., 2010; Sohn et al., 2014).

Earlier medication extraction systems were often rule-based or hybrid. For example, MedEx used semantic rules and regular expressions to extract drug names from clinical narratives (Xu et al., 2010). MedEx-UIMA later extended this line of work by providing an open-source implementation and by mapping drug names to concept codes (Jiang et al., 2014).

A complementary problem is to determine how the drug mention should be interpreted in its clinical context. A drug may be mentioned as a current treatment, a past treatment, a planned prescription, a stopped medication, a hypothetical option, or a negated exposure. This corresponds to drug event qualification rather than simple NER. Recent work and shared tasks have therefore formalized drug extraction as a contextual task, including qualification attributes such as action, negation, temporality, certainty, and actor (Mahajan et al., 2022; Mahajan et al., 2023). This distinction is important for the present chapter: drug name, strength, dosage, and form are treated as entity types in the NER step, whereas action, temporality, certainty, and negation are predicted as qualification attributes.

1.3.3 Laboratory Test Extraction and Normalization

Laboratory test extraction raises a different set of challenges. A laboratory result usually combines several pieces of information: the laboratory test name, a numerical or qualitative value, a unit, and sometimes an interpretation. Extracting this information therefore requires more than detecting a named entity. The system must also identify the associated measurement and standardize it into a usable form. Previous work has studied the extraction of laboratory test information from biomedical or clinical text,

including the association between test names, values, units, and structured laboratory records (Kang et al., 2013; Liu et al., 2017).

Normalization is also particularly important for laboratory tests. Test names may be expressed through abbreviations, local formulations, spelling variants, or incomplete mentions. This makes it difficult to compare mentions across documents or to combine text-derived results with structured data. Biomedical resources and corpora, such as the UMLS provide important support for normalization (Bodenreider, 2004). Recent neural methods for biomedical entity normalization, such as SapBERT and CODER, have improved the ability to match textual mentions with concept codes (Liu et al., 2021a; Yuan et al., 2022c). However, for laboratory-based phenotyping, normalization of the test name must be combined with extraction and standardization of the associated value or interpretation.

1.3.4 Positioning of the Work

The present work builds on these lines of research rather than replacing them. Previous studies have already shown that clinical text can complement structured EHR data, and several systems have addressed drug recognition, normalization, and qualification, as well as laboratory test recognition, measurement extraction, and normalization. The contribution of this chapter is to combine these components within a single applied phenotyping pipeline for French clinical text. We extract and normalize information about drugs and laboratory tests from discharge summaries and apply the resulting pipeline to a real-world cohort from the AP-HP CDW.

From a methodological perspective, we evaluate the different components of the supervised extraction pipeline and further position its NER component with respect to more recent open-vocabulary and instruction-tuned approaches (Zaratiana et al., 2024; Cocchieri et al., 2026; Ge et al., 2026). This provides a contemporary reference point for assessing the performance and computational efficiency of the supervised approach.

More broadly, however, the chapter is not limited to an NLP evaluation. Its main objective is to assess the practical contribution of clinical text to EHR-based phenotyping. Beyond reporting the performance of the individual NLP components, we compare the information available from structured data alone with that obtained after incorporating variables extracted from discharge summaries. This allows us to evaluate whether clinical text provides additional information for characterizing patients with immune-mediated inflammatory diseases.

2 Methods

2.1 Selected Diseases

As a proof of concept, we focused on 4 immune-mediated inflammatory diseases: systemic lupus erythematosus, systemic sclerosis, Antiphospholipid syndrome, and Takayasu arteritis.

2.1.1 Systemic lupus erythematosus

systemic lupus erythematosus (SLE) is an autoimmune disease that mainly affects the skin, joints, and kidneys (Fanouriakis et al., 2019; Aringer et al., 2019). According to the revised 2019 EULAR/ACR classification criteria for SLE, patients are eligible for SLE criteria only if they have a positive antinuclear antibody $\geq 1/80$ at least once. Anti-dsDNA and anti-Smith autoantibodies with high specificity for SLE are also included in the classification criteria for SLE. Therefore, we have chosen these 3 antibodies to identify SLE patients. Hydroxychloroquine, glucocorticoids, mycophenolate mofetil, cyclophosphamide, and belimumab are key treatments of SLE (Haute Autorité de Santé, 2024) and have been chosen to identify patients with SLE.

2.1.2 Systemic sclerosis

Systemic sclerosis is a rare autoimmune disease, inducing skin fibrosis, digestive disorders, such as gastroesophageal reflux disease and chronic pseudoocclusive syndrome, interstitial lung involvement, and sometimes inaugural renal crisis. Classification criteria are also based on specific autoantibodies, including anti-Scl-70, anticentromere, and anti-RNA polymerase III, which we have chosen to analyze here (Haute Autorité de Santé, 2018). Therapeutic management is also based on glucocorticoids and immunosuppressive drugs, such as mycophenolate mofetil.

2.1.3 Antiphospholipid syndrome

Antiphospholipid syndrome (APS) is a systemic autoimmune disease defined by the thrombosis or pregnancy morbidity in the presence of persistent antiphospholipid autoantibodies, lupus anticoagulant, IgG or IgM anticardiolipin, IgG or IgM anti-β2glycoprotein-1 antibodies. Treatment is based on curative anticoagulation with heparin, low-molecular-weight heparin, and antivitamin K (Haute Autorité de Santé, 2022).

2.1.4 Takayasu arteritis

Takayasu arteritis (TA) is an inflammatory disease of the large arteries, leading to arterial stenosis in young people. C-reactive protein is used as an indicator of inflammation and disease activity in TA. Treatment is based on glucocorticoids and immunosuppressive agents such as methotrexate, and biologic agents such as tocilizumab. Therefore, we have chosen to focus our analysis on these treatments.

2.1.5 Vaccinations

Finally, because patients receiving immunosuppressive treatments are at increased risk of infection, vaccination is an important element of their clinical management. In particular, pneumococcal and influenza vaccinations are recommended in this context. We therefore also searched discharge summaries for mentions of these vaccinations.

2.2 Dataset Selection

The dataset used in this study comes from the clinical data warehouse (CDW) of the University Hospitals of Greater Paris (Assistance Publique-Hôpitaux de Paris; AP-HP). The CDW brings together information on all patients followed in the 38 teaching hospitals in the Paris region (>22,000 beds and 1.5 million hospitalizations per year) that use a common EHR software, ORBIS Dedalus Health care. This software has been gradually implemented in the 38 hospitals since 2012.

The dataset was extracted from the CDW research database, in the Informatics for Integrating Biology and the Bedside (i2b2) format (Murphy et al., 2010). The inclusion criteria for the study were as follows: all patients aged >15 years with SLE, systemic sclerosis, APS, or TA who had at least one stay at AP-HP hospitals initially from July 1, 2017, to December, 31, 2020. Patients in the database were selected in 2 ways: by the ICD-10 codes of these 4 pathologies and by keywords present in the medical reports (using regular expression matching), as summarized in Table S1 in Appendix A. For these patients, the data available were demographic data; textual data, including all full-text medical reports, laboratory tests performed during patients' stay, drug prescription, and administration when available; and medico-administrative coding data (ICD-10). The extraction covered all medical departments that could potentially manage patients with the 4 pathologies of interest: internal medicine and clinical immunology, nephrology, rheumatology, dermatology, pneumology, neurology, gastroenterology, oncology, hematology, infectious diseases, and emergency and intensive care.

As this study involves the secondary use of real-life health data, from this large i2b2 extraction, we limited the study to EHRs with at least one ICD-10 code corresponding to the diseases studied (SLE, systemic sclerosis, APS, or TA) and at least one recorded hospital discharge summary, as these are validated by a senior clinician. Subsequently, a subset of this study cohort of 103 hospital discharge summaries, each corresponding to a different patient, was randomly selected and annotated by a clinician (CG), following the same annotation rules as proposed by the national NLP clinical challenges 2022 (Mahajan et al., 2023). Details regarding this annotation process are provided in the section A.2 of Appendix A. The global approach of this work was to build, train, and validate NLP algorithms on the annotated subset before applying it to the full study cohort. Figure 2.2 presents the cohort selection process.

2.3 Knowledge Bases

To combine information extracted from discharge summaries with structured EHR data, mentions detected in text had to be mapped to concept codes. We therefore built one KB for laboratory tests and another for drug treatments, each containing concept codes and their corresponding textual forms.

For drug treatments, the KB was built from two open-source resources. The first was the UMLS (Bodenreider, 2004), restricted to the French ATC vocabulary (World Health Organization, 2024). The second was the Unique Drug Interoperability Repository created by the French National Agency for Medicines and Health Products Safety (Ministère du Travail, de la Santé et des Solidarités et al., 2025). Drug treatments were represented using ATC codes, allowing them to be grouped according to a standard therapeutic classification.

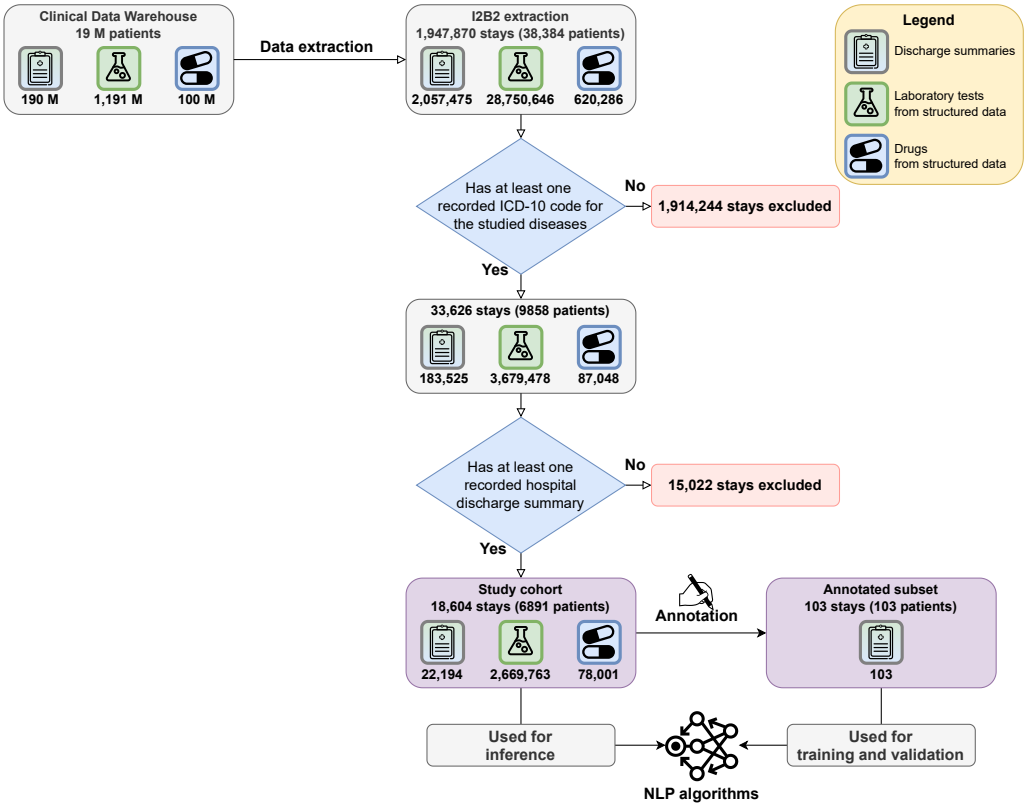


Figure 2.2: Cohort selection flowchart. Starting with integrating biology and bedside extraction of 1,947,870 stays (38,384 patients), the cohort was filtered to include stays with at least 1 International Classification of Diseases, 10th Revision (ICD-10) code corresponding to the studied diseases and at least 1 recorded hospital discharge summary. A final study cohort of 18,604 stays (6891 patients) was created, with 103 randomly selected discharge summaries annotated for training and validation purposes. NLP: natural language processing.

For laboratory tests, the KB was built from the UMLS (Bodenreider, 2004), restricted to the Laboratory Procedure semantic type and to terms from the SNOMED CT US Edition vocabulary (Donnelly, 2006). Both French and English synonyms were retained. Laboratory tests were represented using UMLS Concept Unique Identifiers (CUIs), allowing different textual variants of the same test to be grouped under a common concept code. Table 2.1 summarizes the size of the two KBs used in this chapter.

Entity type	**No. unique concept codes**	**No. corresponding terms**
Laboratory test name	10,370 CUIs	52,240
Drug name	6,707 ATC codes	75,631

Table 2.1: Description of the knowledge bases used for laboratory tests and drug treatments.

2.4 End-to-End Pipeline

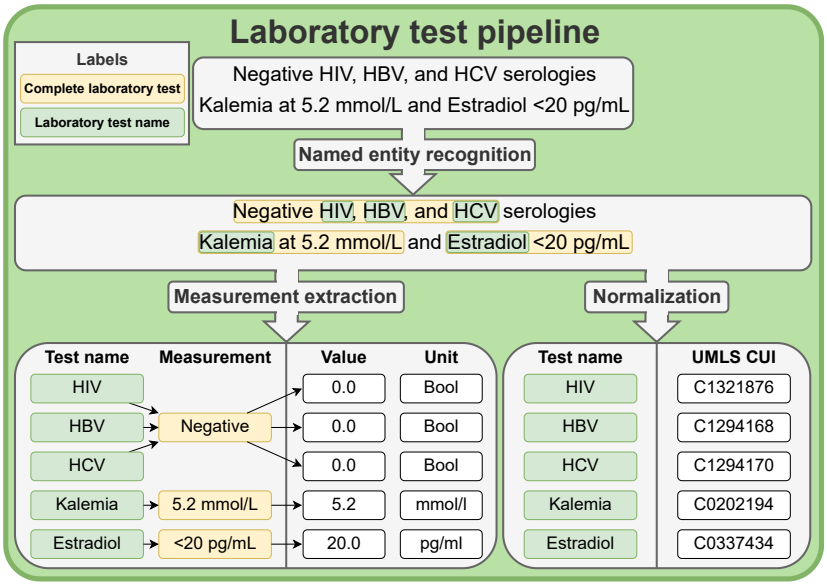
2.4.1 Overview

The system presented in this work required 4 NLP tasks: (1) NER: this task identified and classified entities of interest mentioned in the text into predefined categories. The possible categories included drug name, drug strength, drug dosage, drug form, laboratory test name, and complete laboratory test. (2) Qualification: this task involved assigning predefined qualifiers to the recognized named entities. Only entities classified as "drug name" by the NER algorithm were qualified. There were 4 qualifiers (Mahajan et al., 2022): action (start, stop, increase, decrease, unique dose, and unknown), temporality (present, past, and future), certainty (certain, hypothetical, and conditional), as well as negation (true and false). (3) Measurement extraction: this task extracted and standardized the value and unit contained in the "complete laboratory test" entities detected by the NER algorithm. (4) Normalization: this task assigned predefined concept codes to the recognized named entities. Each entity classified as "drug name" by the NER algorithm was assigned a code from the ATC classification system (World Health Organization, 2024). Each entity classified as "laboratory test name" by the NER algorithm was assigned a CUI of the UMLS (Bodenreider, 2004) restricted to the laboratory procedure semantic type and the SNOMED CT US edition vocabulary (Donnelly, 2006). As described in Figures 2.3a and 2.3b, the laboratory test pipeline and the drug pipeline involved both NER and normalization, while measurement extraction only concerned the laboratory test pipeline and qualification only concerned the drug pipeline.

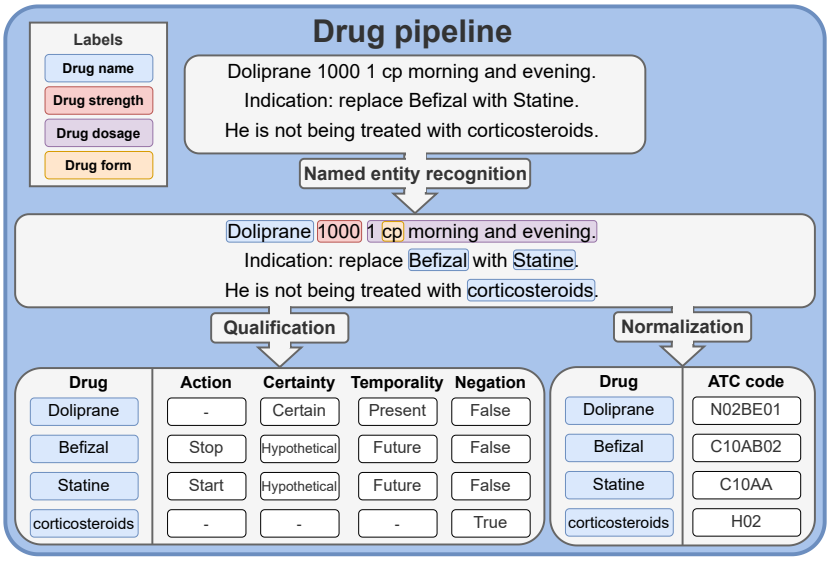
2.4.2 NER and Qualification Tasks

As a preliminary approach for NER, we implemented a rule-based method based on exact matching against terminology dictionaries derived from the target KBs. The ATC classification system (World Health Organization, 2024) was used for drug names, while the SNOMED CT US edition vocabulary (Donnelly, 2006) was used for laboratory tests. A detailed description of these dictionaries is provided in Table 2.1.

We then developed the supervised neural architecture illustrated in Figure 2.4. NER and qualification were modeled as two separate tasks, each using a pretrained BERT



(a) Laboratory test pipeline.



(b) Drug pipeline.

Figure 2.3: Overview of the two NLP pipelines applied to discharge summaries. The laboratory test pipeline extracts and normalizes laboratory tests and their measurements, while the drug pipeline extracts, qualifies, and normalizes drug treatments.

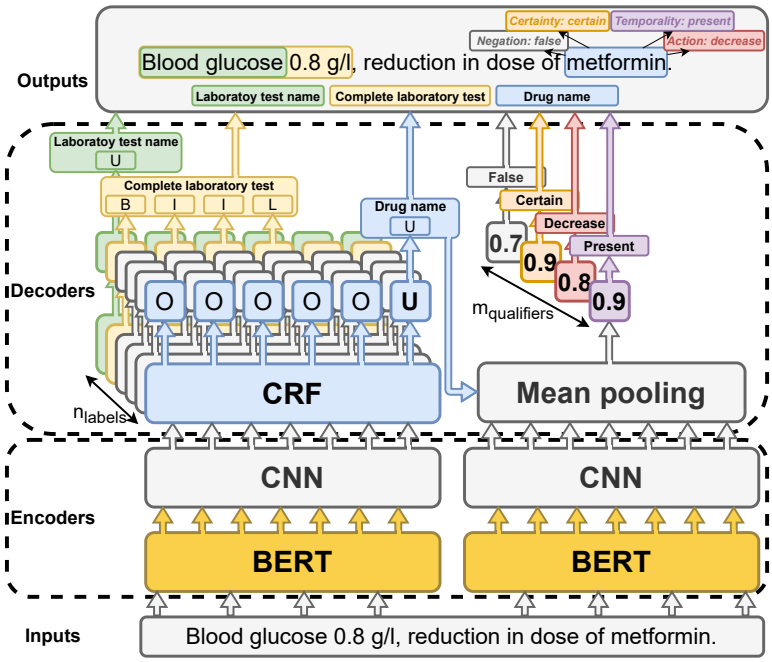


Figure 2.4: Diagram of the named entity recognition and qualification architecture for laboratory test and drug entities. BERT: bidirectional encoder representations from transformers; CNN: convolutional neural network; CRF: conditional random field.

encoder (Devlin et al., 2019). For NER, contextualized token representations are processed by a convolutional neural network (CNN) (Abas et al., 2021) and decoded using a conditional random field (CRF) (Lafferty et al., 2001). Entity spans are represented using the beginning, inside, outside, unit, and last tagging scheme (Ratinov et al., 2009). For qualification, representations associated with each drug entity are processed by a CNN, followed by mean pooling and a softmax classification layer.

Several pretrained French language models were compared as encoders, including CamemBERT-EDS (Dura et al., 2022), CamemBERT-base (Martin et al., 2020), CamemBERT-bio (Touchent et al., 2024), and DrBERT (Labrak et al., 2023). Together with the comparison against the rule-based baseline, these experiments were used to select the supervised configuration retained for subsequent analyses.

We also performed ablation experiments to quantify the contribution of the additional architectural components. For NER, we compared BERT alone, BERT with a CNN, BERT with a CRF, and the complete BERT–CNN–CRF architecture. For qualification, we compared the BERT-based classifier with and without the CNN component.

The supervised models were trained on 83 annotated AP-HP discharge summaries and evaluated on a held-out set of 20 summaries. NER was assessed using precision, recall, and F1-score under strict exact matching, requiring both entity boundaries and labels to match the gold annotation. Qualification was evaluated on gold “drug name” entities. Because qualifier values were highly imbalanced, the most frequent value of each qualifier was treated as the default negative class, and performance was computed for the detection of non-default values. For both tasks, 95% confidence intervals were

Architecture parameters	
Pretrained model	CamemBERT-EDS (Dura et al., 2022)
Transformer window	128 tokens
Transformer stride	96 tokens
CNN kernel size	3
CRF window (NER only)	40 tokens
Fine-tuning hyperparameters	
Batch size	2,000 words
Learning rate	5×10^{-5}
Optimizer	AdamW
Warmup steps	400

Table 2.2: Architecture and fine-tuning configuration of the final supervised NER and qualification models. CNN: convolutional neural network; CRF: conditional random field.

estimated by bootstrap resampling at the discharge-summary level.

To position the selected supervised NER model with respect to more recent approaches, we additionally compared it with open-vocabulary and instruction-tuned methods. We evaluated three GLiNER-based models (Zaratiana et al., 2024): a general-domain model, a biomedical model, and a French biomedical CamemBERT-based model. These models perform span extraction conditioned on natural-language descriptions of the target entity types without task-specific fine-tuning. French descriptive labels were used for the French model and English labels for the English-language models; the complete label mappings are reported in Table S2 in Appendix A. We also evaluated instruction-tuned LLMs using a generative BioNER formulation adapted from Ge et al. (2026), in zero-shot and few-shot settings. For few-shot prompting, examples were dynamically retrieved according to their TF–IDF similarity with the input text. Two LLMs were compared: Qwen3-4B-Instruct, run locally on our computing infrastructure, and GPT-5.6 Luna, accessed through the OpenAI API. The complete prompt and expected output format are provided in Appendix A.4.

The final supervised NER and qualification models used CamemBERT-EDS as the pretrained encoder. Their architecture and fine-tuning configuration are summarized in Table 2.2.

2.4.3 Measurement Extraction Task

Extraction and standardization of the numerical value and unit were applied to the outputs of the NER step, which extracts the complete laboratory test entity from the text in a single block (laboratory test name, numerical value, and unit). The extraction and standardization were achieved with a rule-based algorithm using regular expressions. The algorithm steps are described in Figure 2.5: (1) the laboratory test names were removed from the complete laboratory test entity, (2) regular expressions were designed to extract the numerical or qualitative value and the unit, and (3) qualitative values (eg, “positive," “negative," or “normal") were standardized into graded numbers (1.0, 0.0, or 0.5), while units were converted to conventional standards.

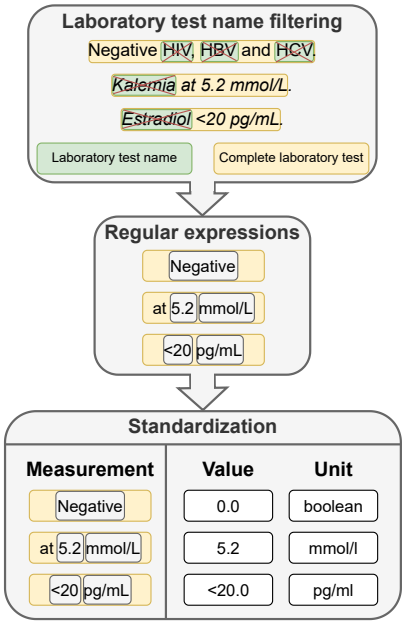


Figure 2.5: Diagram of the laboratory test measurement extraction process. HBV: hepatitis B virus; HCV: hepatitis C virus.

2.4.4 Normalization Task

The aim of the normalization step was to assign a concept code to each detected entity classified as either a “laboratory test name” or a “drug name”. As illustrated in Figure 2.6, each detected mention was compared with the textual forms contained in the corresponding KB introduced in Section 2.3 and summarized in Table 2.1. A similarity score was computed between the mention and each candidate term, and the concept code associated with the highest-scoring term was selected.

We compared two families of similarity methods. The first relied on fuzzy string matching and included Jaro–Winkler distance (Winkler, 1990) and Levenshtein distance (Levenshtein, 1966). The second relied on neural representations, where mentions and candidate terms were embedded in a shared vector space and compared using cosine similarity. We evaluated CODER-all (Yuan et al., 2022c) and SapBERT-all (Liu et al., 2021b).

Normalization was evaluated by comparing the predicted concept code with the annotated gold concept codes. A prediction was considered correct when the predicted code belonged to the set of gold codes associated with the entity. Precision, recall, and F1-score were computed, with 95% confidence intervals estimated by bootstrap resampling at the document level.

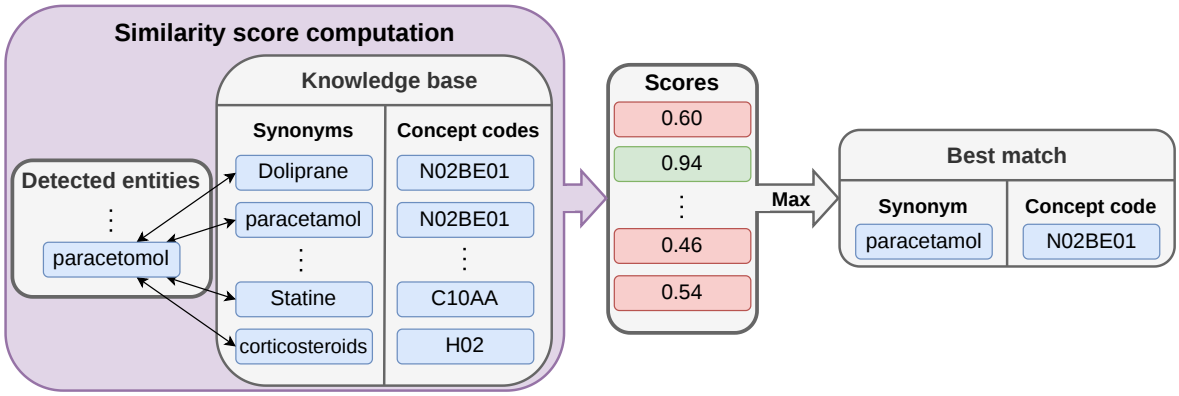


Figure 2.6: Diagram of the normalization process. In this example, the best match of "paracetomol," written with a typo, is "paracetamol," which provides the correct Anatomical Therapeutic Chemical (ATC) code.

2.5 Ethical Considerations

The research protocol for this project was approved in 2020 by the institutional review board of Assistance Publique – Hôpitaux de Paris (AP-HP) (20-93). All data used in this study were collected as part of routine medical care, and their use for research purposes falls under the ethical guidelines of the institutional review board. All patient data were pseudonymized to ensure privacy and comply with data protection regulations. No financial compensation was provided, as the study relied solely on retrospective data from electronic health records.

3 Results

3.1 Dataset Description

The number of patients, hospitalizations, and discharge summaries for each disease in the study cohort are reported in Table 2.3. Figure 2.7 further describes the cohort by showing the age distribution of patients and the distribution of hospitalizations over time for each disease.

We performed various analyses on the extracted data combined with the structured data from the cohort. The results are reported in 2 sections. The first section presents the performance of the NLP algorithms on the annotated subset of 103 discharge summaries. The second section is about the application of these algorithms to the 22,194 discharge summaries included in the study cohort. It described the contribution of unstructured data to structured data.

3.2 NLP Performance

3.2.1 Overview

The performance of the pipeline was evaluated on four main tasks: NER, qualification, measurement extraction, and normalization. This evaluation relied on two datasets: 103

Disease	No. of patients	No. of hospitalizations	No. of discharge summaries
Antiphospholipid syndrome	1,059	1,818	2,380
Lupus	4,102	10,445	12,500
Systemic sclerosis	2,031	6,455	7,585
Takayasu arteritis	252	833	965

Table 2.3: Data description of the study cohort. Number of patients, hospitalizations, and discharge summaries for each studied disease. The number of discharge summaries is higher than the number of hospitalizations, as patients may change departments several times during the same stay (e.g., transfer to an intensive care unit).

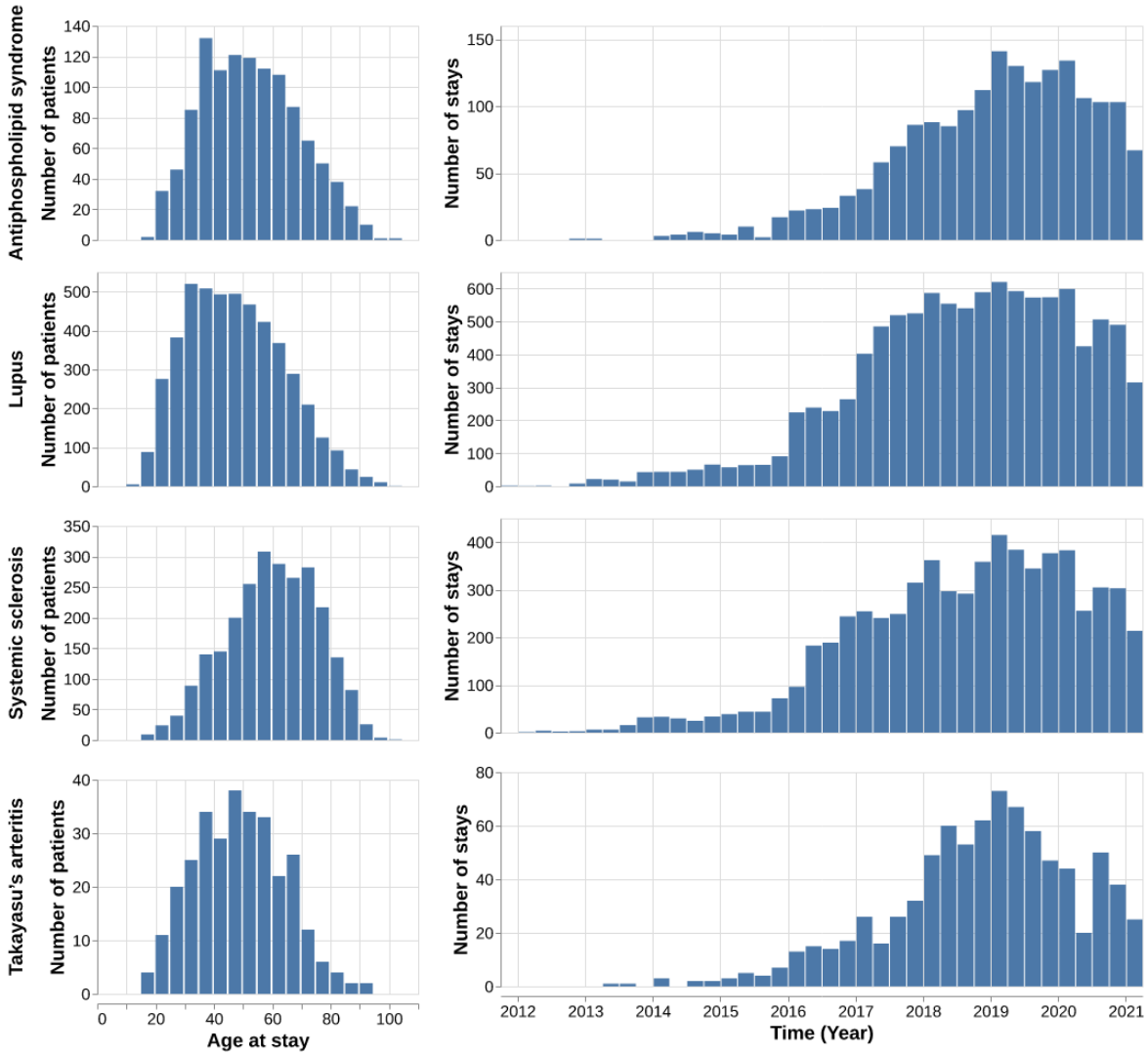


Figure 2.7: Age distributions and hospitalizations over time in the study cohort. The left column shows the age distribution of patients in five-year intervals. The right column shows the number of hospitalizations over time from 2012 to 2021 in quarterly intervals.

manually annotated discharge summaries from the study cohort and the publicly available QUAERO FrenchMed corpus (Névéol et al., 2014). Precision, recall, and F1-score were used as the main evaluation metrics and were reported together with 95% confidence intervals (CIs) estimated using empirical bootstrap resampling at the discharge-summary level (Dekking et al., 2005). The amount of annotated data available for evaluation differed across tasks because the annotation requirements were not identical. NER and qualification required annotated data for both training and testing, whereas measurement extraction and normalization only required annotated examples for evaluation. Consequently, all 103 discharge summaries were annotated for NER and qualification, with 83 documents (80.6%) randomly selected for training and the remaining 20 (19.4%) used as a held-out test set. These 20 test documents were further annotated for measurement extraction. Among them, 11 documents (10.7% of the complete annotated corpus) were additionally annotated for normalization, resulting in 668 normalized entities. This smaller subset reflects the additional annotation effort required to manually assign reference terminology codes to each entity.

3.2.2 NER Task

3.2.2.1 Rule-based We first evaluated the rule-based approach based on exact terminology matching. As shown in Table 2.4, the method achieved an F1-score of 69.3 for drug names and 42.6 for laboratory test names. Performance was particularly limited by recall for laboratory tests, which reached only 33.4, highlighting the difficulty of covering the variability of clinical mentions through exact terminology matching alone.

Label	No. of entities	Precision	Recall	F1-score
Drug name	585 (459–726)	73.8 (65.9–80.4)	65.3 (55.5–73.9)	69.3 (61.1–76.0)
Lab. test name	1,292 (1,009–1,598)	58.5 (53.7–63.1)	33.4 (28.6–38.0)	42.6 (38.0–46.4)
Overall	**1,877 (1,555–2,249)**	**64.8 (61.1–68.7)**	**43.4 (39.8–46.9)**	**52.0 (48.9–54.8)**

Table 2.4: Performance of the rule-based NER approach on the 20 AP-HP discharge summaries of the test set. The approach was evaluated for drug names and laboratory test names, corresponding to the entity types covered by the terminology dictionaries. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals.

3.2.2.2 Pretrained encoder comparison We then compared four pretrained French language models using the same supervised NER configuration to select the encoder. CamemBERT-EDS achieved the highest overall F1-score, at 88.8, followed by CamemBERT-bio at 86.7, CamemBERT-base at 84.4, and DrBERT at 83.2 (Table 2.5). CamemBERT-EDS was therefore retained for the subsequent experiments. Detailed per-entity results for all four pretrained encoders are reported in Tables S3 and S4 in Appendix A.

3.2.2.3 Ablation study After selecting CamemBERT-EDS, we evaluated the contribution of the CNN and CRF components (Table 2.6). The complete BERT–CNN–CRF architecture obtained the highest F1-score, at 88.8. The CRF mainly improved precision, which increased from 80.3 with the encoder alone to 87.5 when the CRF was added. Adding the CNN on top of the CRF provided a smaller additional improvement.

Pretrained encoder	Precision	Recall	F1-score
DrBERT (Labrak et al., 2023)	84.1 (81.6–86.5)	82.3 (79.8–84.5)	83.2 (80.9–85.3)
CamemBERT-base (Martin et al., 2020)	84.2 (81.4–87.0)	84.7 (82.1–87.0)	84.4 (81.9–86.9)
CamemBERT-bio (Touchent et al., 2024)	87.4 (85.3–89.2)	86.1 (83.8–88.0)	86.7 (84.6–88.4)
CamemBERT-EDS (Dura et al., 2022)	**89.1 (87.5–90.8)**	**88.5 (86.4–90.5)**	**88.8 (87.1–90.5)**

Table 2.5: Comparison of pretrained encoders for supervised NER on the AP-HP test set. All models were evaluated using the same NER configuration and the same 20 held-out discharge summaries. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals.

CNN	CRF	Precision	Recall	F1-score
✗	✗	80.3 (80.1–81.8)	87.3 (85.2–89.1)	83.5 (81.8–85.3)
✓	✗	84.4 (83.8–85.7)	88.6 (86.6–90.2)	86.3 (84.7–87.9)
✗	✓	87.5 (86.5–89.0)	**88.9 (87.0–90.6)**	88.2 (86.5–89.7)
✓	✓	**89.1 (87.5–90.8)**	88.5 (86.4–90.5)	**88.8 (87.1–90.5)**

Table 2.6: Ablation study of the CNN and CRF components for NER using CamemBERT-EDS. All configurations were evaluated on the same 20 AP-HP discharge summaries. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals.

3.2.2.4 Final NER performance

The final CamemBERT-EDS–CNN–CRF model achieved an overall F1-score of 88.8, with a precision of 89.1 and a recall of 88.5 (Table 2.7). Across entity types, F1-scores ranged from 84.9 for complete laboratory tests to 93.5 for drug strength.

Label	No. of entities	Precision	Recall	F1-score
Lab. test name	1,292 (1,010–1,603)	90.6 (88.5–92.7)	89.9 (87.5–92.2)	90.3 (88.2–92.3)
Comp. lab. test	1,041 (793–1,323)	86.2 (84.3–88.3)	83.6 (79.1–87.5)	84.9 (82.1–87.5)
Drug name	585 (454–731)	90.8 (85.8–95.4)	92.5 (88.4–95.7)	91.6 (87.5–95.1)
Drug dosage	276 (194–368)	89.1 (85.4–94.1)	86.2 (81.7–90.8)	87.7 (84.0–92.1)
Drug form	170 (106–247)	86.5 (78.8–93.4)	94.1 (91.5–97.6)	90.1 (85.6–94.0)
Drug strength	130 (77–196)	93.1 (87.7–97.0)	93.8 (89.2–97.6)	93.5 (89.3–96.6)
Overall	**3,494 (2,885–4,194)**	**89.1 (87.5–90.8)**	**88.5 (86.4–90.5)**	**88.8 (87.1–90.5)**

Table 2.7: Performance of the final CamemBERT-EDS–CNN–CRF model for NER on the 20 AP-HP discharge summaries of the test set. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals.

3.2.2.5 Impact of training set size

We also assessed the influence of the amount of annotated training data on the AP-HP NER task. As shown in Figure 2.8, performance increased rapidly with the first annotated discharge summaries and then progressively approached a plateau after approximately 20 training documents. Additional annotations continued to improve performance, but with smaller gains.

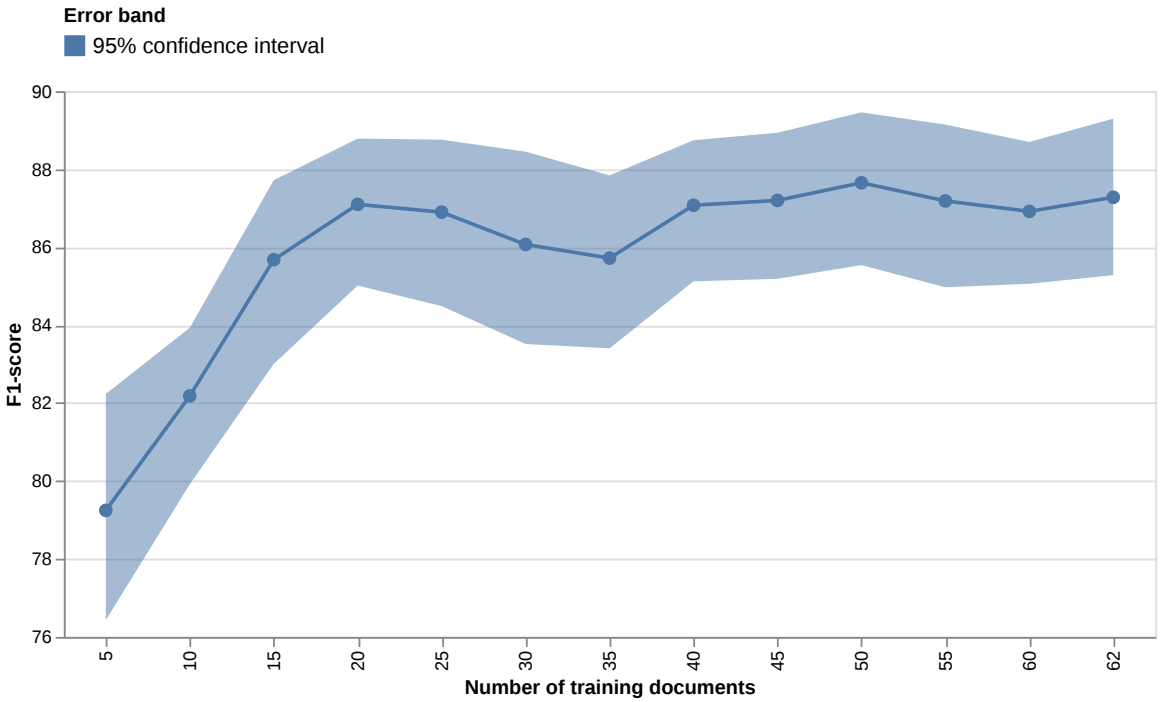


Figure 2.8: Evolution of NER performance according to the number of annotated AP-HP discharge summaries used for training. The curve shows the F1-score, with the shaded area representing the 95% confidence interval.

3.2.2.6 Comparison with recent NER approaches While the preceding experiments were conducted on AP-HP discharge summaries, we additionally used the QUAERO FrenchMed corpus to position the selected supervised model with respect to more recent open-vocabulary and instruction-tuned approaches. Results are reported separately for the EMEA and MEDLINE subsets in Table 2.8, while detailed per-entity results for the supervised CamemBERT-EDS model are provided in Tables S5 and S6 in Appendix A. On QUAERO EMEA, CamemBERT-EDS achieved the highest F1-score, at 71.9, compared with the best GLiNER-based and instruction-tuned LLM configurations. On QUAERO MEDLINE, CamemBERT-EDS achieved an F1-score of 64.2. The supervised model therefore remained the strongest approach in terms of F1-score on both subsets. The GLiNER-based approaches offered the advantage of predicting entities directly from natural-language descriptions of the target types, but their F1-scores remained below those of CamemBERT-EDS on both QUAERO subsets. For the instruction-tuned LLMs, few-shot prompting improved performance over the corresponding zero-shot configurations, showing the benefit of providing examples selected from the training data. Nevertheless, the supervised CamemBERT-EDS model remained the highest-performing approach overall. The inference-time results favored the encoder-based approaches, with both CamemBERT-EDS and the GLiNER models being substantially faster than the instruction-tuned LLMs. The GPT-5.6 Luna timings should nevertheless be interpreted cautiously, as the model was accessed through the OpenAI API, ran on different infrastructure, and the reported time also included network communication.

Method	Precision	Recall	F1-score	sec/doc
QUAERO EMEA				
CamemBERT-EDS	79.9 (77.1–82.7)	65.3 (59.7–70.8)	**71.9 (67.8–75.6)**	0.436
CamemBERT-GLiNER	44.9 (41.9–48.4)	25.7 (22.8–29.4)	32.7 (29.8–36.1)	0.154
GLiNER-BioMed	53.1 (50.2–56.4)	53.1 (48.9–57.1)	53.1 (49.6–56.5)	0.260
GLiNER	61.9 (58.6–65.3)	47.9 (41.9–53.5)	54.0 (49.2–58.1)	0.246
Qwen3-4B (zero-shot)	26.5 (18.9–35.1)	10.3 (9.3–11.8)	14.8 (13.3–17.0)	37.07
Qwen3-4B (few-shot)	43.4 (41.0–46.0)	39.5 (35.6–43.5)	41.4 (38.3–44.4)	84.61
GPT-5.6 Luna (zero-shot)	52.3 (48.7–55.7)	51.1 (45.9–56.5)	51.7 (47.4–55.9)	20.39
GPT-5.6 Luna (few-shot)	60.0 (57.0–62.9)	57.0 (51.1–62.9)	58.5 (54.1–62.7)	18.14
QUAERO MEDLINE				
CamemBERT-EDS	72.1 (70.2–73.9)	57.9 (56.3–59.5)	**64.2 (62.5–65.8)**	0.003
CamemBERT-GLiNER	40.0 (38.4–41.5)	49.5 (47.8–50.9)	44.2 (42.7–45.7)	0.030
GLiNER-BioMed	33.4 (32.0–34.6)	69.8 (68.2–71.5)	45.2 (43.7–46.5)	0.031
GLiNER	36.0 (34.5–37.2)	65.1 (63.5–66.8)	46.4 (44.8–47.6)	0.033
Qwen3-4B (zero-shot)	49.1 (46.4–51.3)	35.9 (34.5–37.4)	41.5 (39.6–43.2)	1.98
Qwen3-4B (few-shot)	51.6 (49.6–53.3)	56.6 (55.0–58.3)	54.0 (52.3–55.5)	2.37
GPT-5.6 Luna (zero-shot)	53.2 (51.2–54.8)	62.4 (60.9–63.9)	57.4 (55.8–58.9)	2.25
GPT-5.6 Luna (few-shot)	57.4 (55.5–59.2)	65.5 (63.6–67.3)	61.2 (59.5–62.7)	1.05

Table 2.8: Comparison of supervised, open-vocabulary, and instruction-tuned NER approaches on the QUAERO EMEA and MEDLINE corpora. Precision, recall, and F1-score were computed using strict exact matching and are reported with document-level bootstrap 95% confidence intervals. Few-shot LLM demonstrations were retrieved according to TF–IDF similarity with the input document. Qwen3-4B-Instruct was run locally, whereas GPT-5.6 Luna was accessed through the OpenAI API. Inference time is reported in seconds per document; API-based timings include communication and network latency and are therefore not directly comparable to local inference times.

3.2.3 Qualification Task

The qualification experiments were conducted on the AP-HP annotated discharge summaries, using the same split of 83 documents for training and 20 held-out documents for evaluation.

3.2.3.1 Ablation study We first evaluated the contribution of the CNN component to the qualification model (Table 2.9). Without the CNN, the model achieved an F1-score of 78.1, with a precision of 78.4 and a recall of 77.8. Adding the CNN slightly improved all three metrics, resulting in an F1-score of 78.8, with a precision of 79.1 and a recall of 78.6. The CNN-based configuration was therefore retained for the final qualification model.

3.2.3.2 Final qualification performance The final qualification model achieved an overall F1-score of 78.8, with a precision of 79.1 and a recall of 78.6 (Table 2.10). Performance varied across qualifiers, with the highest F1-score obtained for negation (85.4), followed by temporality (80.6) and action (77.4). Certainty was the most difficult qualifier, with an F1-score of 69.6 and a wider confidence interval, consistent with the smaller number of non-default examples available for this qualifier.

CNN	Precision	Recall	F1-score
✗	78.4 (72.5–83.4)	77.8 (70.9–83.6)	78.1 (72.4–82.9)
✓	**79.1 (73.3–84.0)**	**78.6 (71.8–84.3)**	**78.8 (73.2–83.6)**

Table 2.9: Ablation study of the CNN component for the qualification task on the AP-HP discharge summaries. Both configurations were evaluated on the same 20 held-out discharge summaries. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals.

Qualifier	No. of entities	Precision	Recall	F1-score
Action	179 (111–267)	78.3 (66.4–87.1)	76.5 (69.0–82.5)	77.4 (68.5–83.4)
Certainty	53 (26–85)	82.1 (56.8–100.0)	60.4 (39.5–82.1)	69.6 (50.0–86.2)
Negation	42 (27–56)	87.5 (76.3–95.7)	83.3 (69.4–97.2)	85.4 (75.0–94.3)
Temporality	216 (144–303)	77.7 (68.7–84.8)	83.8 (74.0–91.5)	80.6 (72.7–86.3)
Overall	**490 (348–661)**	**79.1 (73.3–84.0)**	**78.6 (71.8–84.3)**	**78.8 (73.2–83.6)**

Table 2.10: Performance of the final qualification model on the 20 held-out AP-HP discharge summaries. Precision, recall, and F1-score are reported for non-default qualification values, with document-level bootstrap 95% confidence intervals.

3.2.4 Measurement Extraction Task

The model was evaluated for the measurement extraction task on 20 annotated discharge summaries from AP-HP. Precision, recall, and F1-score are presented in Table 2.11 and were measured in a "strict" way: a true positive result was obtained when a predicted measurement and a gold measurement had the same value and unit. The algorithm obtained an F1-score of 96.7.

No. of entities (95% CI)	Precision (95% CI)	Recall (95% CI)	F1-score (95% CI)
789 (557–1038)	98.2 (96.9–99.3)	95.2 (91.2–97.8)	96.7 (94.5–98.1)

Table 2.11: Performance of the model for measurement extraction. This rule-based algorithm was evaluated on 20 annotated discharge summaries. Each result was bootstrapped by discharge summary to provide a 95% CI.

3.2.5 Normalization Task

3.2.5.1 Comparison of normalization methods We first compared rule-based and embedding-based normalization methods on 11 annotated AP-HP discharge summaries (Table 2.12). For laboratory test names, CODER-all achieved the highest F1-score, at 82.2, compared with 62.6 for the best rule-based approach, Jaro–Winkler distance. CODER-all was therefore retained for laboratory test normalization. For drug names, Jaro–Winkler distance achieved the highest F1-score, at 92.9, compared with 89.7 for the best embedding-based approach, CODER-all. Given its strong performance and lower computational cost, Jaro–Winkler distance was retained for drug normalization.

Label	Method	Precision	Recall	F1-score
Lab. test name	Jaro–Winkler	71.5 (62.2–76.7)	55.6 (46.0–62.5)	62.6 (53.0–68.8)
	Levenshtein	61.6 (51.8–67.6)	58.1 (49.3–66.3)	59.8 (50.4–66.8)
	CODER-all	**92.6 (89.8–95.6)**	**73.9 (68.4–79.7)**	**82.2 (77.9–86.4)**
	SapBERT-all	73.3 (68.7–77.0)	73.3 (68.7–77.0)	73.3 (68.7–77.0)
Drug name	Jaro–Winkler	95.9 (93.8–97.6)	**90.1 (86.9–92.5)**	**92.9 (90.3–94.8)**
	Levenshtein	92.1 (89.4–94.3)	89.7 (87.2–91.9)	90.9 (88.3–93.0)
	CODER-all	**96.7 (94.1–98.9)**	83.7 (79.8–87.3)	89.7 (86.8–92.3)
	SapBERT-all	84.0 (79.7–88.2)	84.0 (79.7–88.2)	84.0 (79.7–88.2)

Table 2.12: Comparison of normalization methods on 11 annotated AP-HP discharge summaries. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals.

3.2.5.2 Final normalization performance The selected normalization methods were then evaluated on the 11 annotated AP-HP discharge summaries and on the EMEA and MEDLINE subsets of the QUAERO FrenchMed corpus (Névéol et al., 2014). CODER-all was used for laboratory test normalization, while Jaro–Winkler distance was used for drug normalization. Results are reported in Table 2.13.

On the AP-HP discharge summaries, laboratory test normalization achieved an F1-score of 82.2, while drug normalization reached 92.9. On QUAERO EMEA, F1-scores were 87.2 for laboratory tests and 88.6 for drugs. On QUAERO MEDLINE, the corresponding F1-scores were 71.1 and 84.7.

Label	Dataset	No. of entities	Precision	Recall	F1-score
Lab. test name	Discharge summaries	356 (199–551)	92.6 (89.7–95.6)	73.9 (68.6–79.6)	82.2 (77.8–86.4)
	EMEA	20 (12–30)	89.5 (60.0–100.0)	85.0 (53.3–100.0)	87.2 (57.9–100.0)
	MEDLINE	81 (74–88)	88.9 (80.3–96.4)	59.3 (48.1–70.6)	71.1 (60.6–80.6)
Drug name	Discharge summaries	312 (204–425)	95.9 (93.9–97.7)	90.1 (86.6–92.5)	92.9 (90.2–94.8)
	EMEA	734 (576–918)	92.2 (89.8–94.6)	85.3 (79.9–89.8)	88.6 (85.2–91.6)
	MEDLINE	334 (318–350)	88.3 (84.6–91.9)	81.4 (77.0–85.6)	84.7 (80.7–88.5)

Table 2.13: Performance of the selected normalization methods on 11 annotated AP-HP discharge summaries and on the QUAERO FrenchMed corpus. CODER-all was used for laboratory test normalization and Jaro–Winkler distance for drug normalization. The QUAERO evaluation included 3 EMEA documents and 833 MEDLINE titles. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals.

3.2.6 End-to-End Pipeline

The model was evaluated on the NER and normalization task using 11 annotated discharge summaries from AP-HP. Precision, recall, and F1-score were measured in a

"strict" way: a true positive result was obtained when a predicted entity and a gold entity had exactly the same boundaries and label and the predicted concept code of the entity was part of the list of annotated gold concept codes. Precision, recall, and F1-score are presented in Table 2.14. We obtained an F1-score of 71.1 for laboratory tests and 89.3 for drug names.

Label	No. of entities (95% CI)	Precision (95% CI)	Recall (95% CI)	F1-score (95% CI)
Lab. test name	356 (204–548)	72.0 (65.3–79.1)	70.2 (59.7–77.9)	71.1 (63.6–77.8)
Drug name	312 (206–424)	91.9 (88.9–93.5)	86.9 (82.9–90.0)	89.3 (85.9–91.6)

Table 2.14: Performance of the models for the combined named entity recognition and normalization tasks on AP-HP CDW discharge summaries. The model was tested on 11 discharge summaries. Each result was bootstrapped by discharge summary to provide a 95% confidence interval (CI), shown in parentheses.

3.3 Clinical Application

After evaluating the intrinsic performance of the NLP pipeline, we applied it to the full study cohort to assess its contribution to patient phenotyping. For each studied disease (SLE, systemic sclerosis, APS, and TA), each studied antibody, and each studied drug treatment, we reported the number of patients for whom we extracted a positive antibody or a prescribed drug treatment from both the structured and unstructured data of the study cohort. Lists of CUI codes for the studied antibodies and ATC codes for drug treatments are available in Tables S7 and S8 in Appendix A. We were particularly interested in the number of patients for whom we extracted positive antibodies or drug treatments from the unstructured data of their EHRs that were not available in the structured data. In the analysis of the unstructured data, only entities explicitly qualified as "certain" and not negated were retained.

3.3.1 Laboratory Test Results

Autoantibody positivity.

Table 2.15 reports the number of antibody-positive patients for each disease, where positivity was determined either by test values exceeding reference ranges or by explicit interpretation of the test as positive. The first column shows the number of patients for whom we extracted at least 1 positive test from the structured data of their EHR. The second column shows the number of patients for whom we extracted at least 1 positive test from both the structured and unstructured data. The third column shows the number of patients for whom we extracted at least 1 positive test from the unstructured data, but only among patients without positive tests extracted from the structured data.

To ensure the reliability of the detected autoantibodies across the entire study cohort, we conducted a second evaluation focused specifically on the studied autoantibodies. While the precision was reported in Table 2.14 as 72% in the general case for all laboratory tests, this additional analysis aimed to ensure comparable performance for the autoantibodies studied. For this evaluation, we randomly selected 10 positively detected entities for each studied autoantibody, yielding a total of 110 entities. These entities,

Disease and lab. test	Structured data only, n (%)	Structured and unstructured data, n (%)	Benefit of unstructured data, n (%)
Antiphospholipid syndrome (1,059 patients)			
Anticardiolipin antibody	184 (17.37)	478 (45.14)	294 (27.76)
Anti-B2GP1 antibody	103 (9.73)	334 (31.54)	231 (21.81)
Lupus anticoagulant	277 (26.16)	423 (39.94)	146 (13.79)
Systemic lupus erythematosus (4,102 patients)			
Antinuclear antibody	752 (18.33)	2,949 (71.89)	2,197 (53.56)
Anti-DNA antibodies	541 (13.19)	2,174 (53.00)	1,633 (39.81)
Anti-Smith antibodies	255 (6.22)	858 (20.92)	603 (14.70)
Systemic sclerosis (2,031 patients)			
Anti-RNA polymerase III antibody	39 (1.92)	156 (7.68)	117 (5.76)
Anti-Scl-70 antibodies	97 (4.78)	568 (27.97)	471 (23.19)
Anticentromere antibody	144 (7.09)	609 (29.99)	465 (22.90)

Table 2.15: Number of patients with positive antibodies identified from structured data and from the combination of structured and unstructured data. The last column reports the additional number of patients identified only through unstructured data.

identified through NER, measurement extraction, and normalization, were reviewed by a clinician. The review identified 3 errors, resulting in a precision of 97.3%.

We also examined EHR cases in which a positive autoantibody appeared in the structured data but was not detected in the unstructured data. For each antibody, we randomly selected 10 EHRs; for some antibodies, fewer than 10 documents met this criterion, so we included all available cases. A clinician reviewed 63 discharge summaries from the EHRs and identified 2 types of situations. In 23 (26.5%) summaries, the algorithm either failed to detect the relevant antibody or did not normalize it correctly. In the remaining 40 (63.5%) summaries, either the antibody was not mentioned in the text, or the mention was interpreted as a negative result by the clinician.

Quantitative laboratory measurements.

Beyond autoantibody positivity, we examined whether the pipeline could also recover common quantitative laboratory measurements from discharge summaries. Figure 2.9 compares the distributions of creatinine, hemoglobin, C-reactive protein (CRP), international normalized ratio (INR), and glomerular filtration rate (GFR) values obtained from structured data, from additional measurements extracted from unstructured data, and from the combination of both sources. This analysis was performed on the whole study cohort, pooling the 4 studied diseases.

For creatinine, hemoglobin, and CRP, the additional text-derived measurements followed distributions visually similar to those observed in structured data. Adding discharge summaries increased the number of available measurements from 31,859 to 40,871 for creatinine, from 29,977 to 39,116 for hemoglobin, and from 20,283 to 28,723 for CRP. It also increased patient coverage from 65.8% to 90.1% for creatinine, from 63.6% to 92.2% for hemoglobin, and from 61.5% to 86.7% for CRP. For INR, the number of measurements increased from 6,541 to 9,229, and patient coverage increased from 24.7% to 35.5%, although the distribution of text-derived values appeared less similar to that of structured values. Finally, no additional GFR value was recovered from unstructured data, and the number of available measurements and covered patients remained unchanged at 12,097 measurements and 2,041 patients (29.6%).

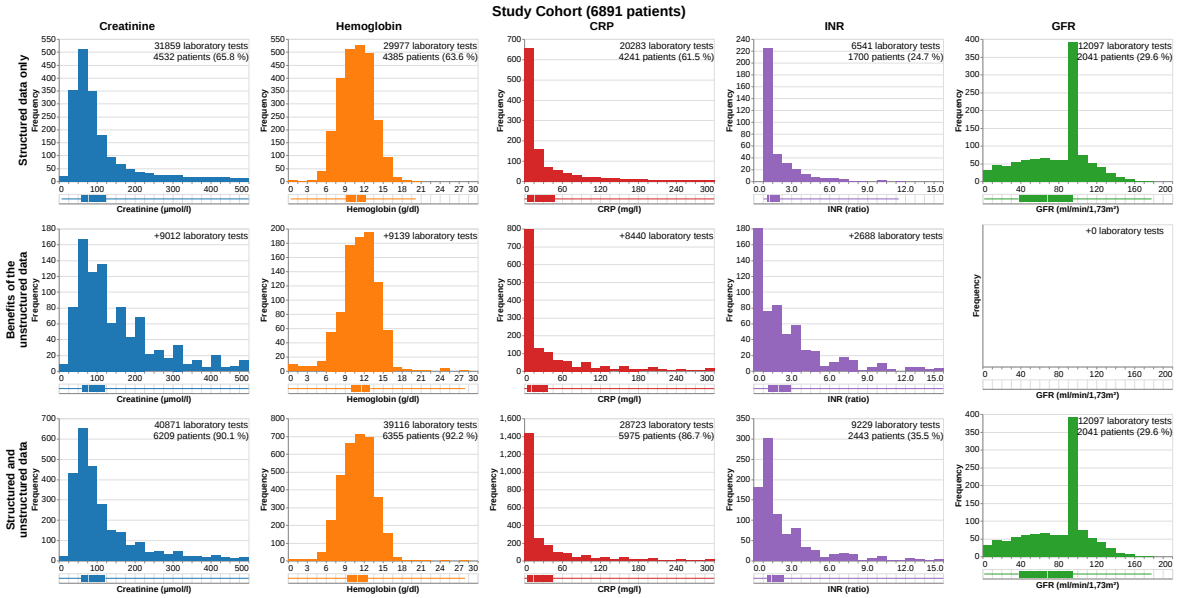


Figure 2.9: Distribution of selected quantitative laboratory test results in the study cohort. For each laboratory test, the first row shows values available from structured data only, the second row shows the additional values extracted from unstructured discharge summaries, and the third row shows the combination of structured and unstructured data. The analysis was performed on the whole study cohort, pooling the 4 studied diseases. CRP: C-reactive protein; INR: international normalized ratio; GFR: glomerular filtration rate.

3.3.2 Drug Treatments

Table 2.16 reports the number of patients with drug treatments for each of the studied drugs related to the respective disease. The first column presents the number of patients for whom we extracted at least 1 drug treatment from the structured data of their EHR. The second column presents the number of patients for whom we extracted at least 1 drug treatment from both the structured and the unstructured data. The third column provides the number of patients for whom we extracted at least 1 drug treatment from the unstructured data but only among the patients without drug treatment extracted from the structured data.

To ensure the precision of the drug treatments identified by the algorithm, we conducted a second evaluation specifically focused on the studied drugs. Although the general precision for all drugs had previously been reported as 91.9% in Table 2.14, this additional analysis aimed to confirm comparable performance for the specific drug treatments studied. For this evaluation, we randomly selected 10 positively detected entities per studied drug (spanning NER and normalization), resulting in a total of 130 entities. A clinician reviewed these entities individually and found no errors, corresponding to a precision of 100%.

3.3.3 Inference Time and Carbon Footprint

The scalability of the pipeline is an important point when considering its use in a CDW. The full NLP process was applied to 22,194 discharge summaries. Inference took 145

Disease and drug	Structured data only, n (%)	Structured and unstructured data, n (%)	Benefit of unstructured data, n (%)
Antiphospholipid syndrome (1,059 patients)			
VKA	186 (17.56)	628 (59.30)	442 (41.74)
Heparin	238 (22.47)	677 (63.94)	439 (41.46)
Oral anticoagulant	47 (4.44)	177 (16.72)	130 (12.28)
Systemic lupus erythematosus (4,102 patients)			
Systemic glucocorticoids	950 (23.16)	3,308 (80.64)	2,358 (57.49)
Cyclophosphamide	64 (1.56)	894 (21.79)	830 (20.23)
Mycophenolate mofetil	301 (7.34)	1,263 (30.78)	962 (23.46)
Rituximab	75 (1.83)	709 (17.28)	634 (15.46)
Belimumab	43 (1.05)	247 (6.02)	204 (4.97)
Methotrexate	112 (2.73)	963 (23.48)	851 (20.75)
Hydroxychloroquine	920 (22.43)	3,520 (85.83)	2,600 (63.40)
Prevenar 13 vaccine	122 (2.97)	984 (23.99)	862 (21.02)
Pneumovax vaccine	43 (1.05)	436 (10.63)	393 (9.58)
Influenza vaccine	96 (2.34)	577 (14.07)	481 (11.73)
Systemic sclerosis (2,031 patients)			
Systemic glucocorticoids	258 (12.71)	1,260 (62.06)	1,002 (49.33)
Cyclophosphamide	6 (0.30)	390 (19.20)	384 (18.91)
Mycophenolate mofetil	88 (4.33)	463 (22.81)	375 (18.47)
Rituximab	13 (0.64)	258 (12.71)	245 (12.07)
Methotrexate	66 (3.25)	541 (26.63)	475 (23.39)
Prevenar 13 vaccine	59 (2.90)	545 (26.84)	486 (23.93)
Pneumovax vaccine	13 (0.64)	285 (14.03)	272 (13.40)
Influenza vaccine	42 (2.07)	425 (20.93)	383 (18.87)
Takayasu arteritis (252 patients)			
Systemic glucocorticoids	68 (27.00)	223 (88.49)	155 (61.51)
Cyclophosphamide	0 (0.00)	18 (7.14)	18 (7.14)
Tocilizumab	15 (5.95)	47 (18.65)	32 (12.70)
Mycophenolate mofetil	9 (3.57)	21 (8.33)	12 (4.76)
Rituximab	0 (0.00)	6 (2.38)	6 (2.38)
Methotrexate	20 (7.94)	135 (53.57)	115 (45.63)
Prevenar 13 vaccine	9 (3.57)	83 (32.94)	74 (29.37)
Pneumovax vaccine	3 (1.19)	48 (19.05)	45 (17.86)
Influenza vaccine	6 (2.38)	41 (16.27)	35 (13.89)

Table 2.16: Number of patients with drug treatments identified from structured data and from the combination of structured and unstructured data. The last column reports the additional number of patients identified only through unstructured data.

minutes on a Tesla V100-SXM2-32GB graphics processing unit and produced an estimated emission of 0.39 kg of CO_2 equivalent. These results suggest that the pipeline can be applied at the scale of a large clinical cohort, although deployment on continuously updated CDW data would require additional engineering.

4 Discussion

4.1 Main Findings

In this chapter, we proposed an end-to-end information extraction pipeline designed to enrich structured EHR data with information extracted from French discharge summaries. The pipeline combines four successive steps: named entity recognition, qualifi-

cation, measurement extraction, and normalization. It was evaluated both intrinsically, using manually annotated discharge summaries and the publicly available Quaero corpus (Névéol et al., 2014), and extrinsically, by measuring its contribution to patient phenotyping in a large AP-HP cohort. In addition, several models and methods were compared for the different steps of the pipeline, making it possible to evaluate not only the final system, but also the methodological choices behind it.

Overall, the pipeline achieved strong performance on the annotated AP-HP discharge summaries. The best NER model obtained high F1-scores for both laboratory test names and drug names, and the qualification and measurement extraction modules also performed well. These results support the use of BERT-based models for clinical NER, a task for which such architectures remain strong and widely used baselines (Chen et al., 2025b). However, the end-to-end results also showed a clear difference between entity types: performance was higher for drug treatments than for laboratory tests. This difference suggests that detecting clinical mentions is only one part of the problem, normalizing these mentions into reusable concept codes is also a major step, especially for laboratory data.

At the cohort level, adding information extracted from discharge summaries substantially increased the availability of clinically relevant variables. This was observed for autoantibody positivity and drug treatments, but also for several quantitative laboratory measurements. For creatinine, hemoglobin, and CRP, the additional values extracted from text increased patient coverage and followed distributions broadly consistent with those observed in structured data. In contrast, INR values extracted from text appeared more selected, and no additional GFR values were recovered. These results illustrate both the value and the variability of clinical text: discharge summaries can provide information that is missing from structured tables, but their contribution depends on the variable considered and on the ability of the pipeline to normalize it correctly.

4.2 Clinical Relevance

The extracted variables were clinically plausible and consistent with previous knowledge about the studied diseases. For example, the proportion of patients with SLE and positive antinuclear antibodies was consistent with the expected clinical profile of the disease (Haute Autorité de Santé, 2024). Similarly, the proportions of patients with SLE treated with hydroxychloroquine, corticosteroids, or immunosuppressants were close to those reported in previous studies and real-world data (Fanouriakis et al., 2019; Eviatar et al., 2024). The results were also coherent for the other diseases: antibody positivity was frequent in APS, and the treatments identified in patients with Takayasu's arteritis, including systemic corticosteroids and tocilizumab, were consistent with national recommendations (Haute Autorité de Santé, 2020).

This type of enriched phenotyping has several potential applications. In research, it can support more precise cohort construction and facilitate the identification of patients eligible for observational studies or therapeutic trials. In clinical research settings, it may also help compare patients with similar immune profiles, for example patients sharing the same autoantibody pattern or treatment history. More detailed analyses could then study the relationship between antibody profiles, organ damage, and treatment trajectories. Examples include the association between anti-RNA polymerase III antibodies and scleroderma renal crisis (Mouthon et al., 2014), or the prognostic value of

triple positivity in APS. Treatment dosages and temporal treatment changes could also be integrated in future analyses.

However, the results also show that clinical text should be seen as complementary to structured data, not as a replacement. Structured data are often more systematic for some laboratory measurements, whereas discharge summaries contain selected, interpreted, and clinically synthesized information. The a posteriori validation supports the reliability of the extracted variables for the targeted downstream analyses, with a precision of 97.3% for the studied autoantibodies and 100% for the reviewed drug entities. Still, the strongest phenotyping strategy is to combine structured and unstructured sources, because each captures different aspects of the patient record.

4.3 Limitations

This study has several limitations. First, the end-to-end evaluation was performed on a limited number of annotated documents: 11 discharge summaries, containing 668 annotated entities. This limitation is mainly due to the cost of manual annotation, especially for normalization, where the annotator must not only identify a mention but also assign the correct concept code, such as a CUI or an ATC code. In addition, the annotations were produced by a single expert, so inter-annotator agreement could not be computed. Several precautions were taken to reduce bias: the annotations followed detailed guidelines, the annotator was an expert clinician, the model was developed independently, and the training and test sets contained discharge summaries from distinct patients. Nevertheless, the limited size of the annotated evaluation set should be kept in mind when interpreting the results.

Beyond the size of the evaluation set, the need for annotation is itself an important limitation of the approach. The initial rule-based method, which relied directly on existing terminologies, did not reach the performance of the supervised models. This suggests that terminologies alone are not sufficient to extract clinically useful variables from discharge summaries with high accuracy. Manual annotation was therefore not only required to evaluate the system, but also to define the target information precisely and to train models adapted to the writing style of AP-HP discharge summaries. This limits the immediate scalability of the approach to new clinical variables or new document types, for which additional annotation and validation would likely be needed.

Second, this work relies on the secondary use of routinely collected hospital data. Such data are rich, but they are not collected primarily for research. Discharge summaries do not contain all patient characteristics, all laboratory results, or all treatment details. This was illustrated by the error analysis of cases where a positive biological test was present in structured data but not detected in text: in many cases, the test was either not mentioned in the discharge summary or was interpreted as negative by the clinician. Therefore, absence of information in text should not be interpreted as absence of the corresponding clinical fact. This reinforces the need to combine structured data and clinical narratives when building research-ready phenotypes.

A third limitation concerns normalization. Drug names are often written in relatively standardized forms, using either trade names or generic names. However, abbreviated drug mentions can still be difficult to normalize. For example, antivitamin K, one of the treatment studied in the antiphospholipid syndrome cohort, was sometimes written as “AVK”, which was not correctly normalized by the algorithm. Laboratory tests are

even more variable. They can be written using abbreviations, local variants, indirect formulations, or values without a clear canonical test name. For example, hemoglobin may appear as “hemoglobin”, “Hb”, or indirectly through expressions such as “anemia at 9 g/dL”. Similarly, abbreviations such as “ACC” for lupus anticoagulant or “DFG” for glomerular filtration rate. In the present pipeline, normalization mainly relied on the surface form of the detected mention, using dictionary matching or similarity-based methods. This can fail when the correct concept depends on the surrounding context. This limitation motivates the next chapter, which focuses specifically on biomedical entity normalization and investigates how contextual training examples can be generated when manually annotated data are scarce.

5 Conclusion

This chapter demonstrated the feasibility of extracting and normalizing clinically relevant information from French discharge summaries at the scale of a large hospital cohort. By combining NER, qualification, measurement extraction, and normalization, the proposed pipeline enriched structured EHR data with information on laboratory tests and drug treatments. To our knowledge, this work is among the first to automatically characterize such a large cohort of patients with autoimmune diseases using information extracted directly from clinical text.

The results show that unstructured clinical text can recover meaningful information that is not fully captured in structured tables. They also show that text-derived and structured data are complementary: some information is found in discharge summaries, while other information remains available only in structured EHR data. These findings should nevertheless be interpreted within the scope of this proof of concept, which focused on a specific cohort, two types of information, and one main document type. Its generalization to other clinical variables, patient populations, or document types therefore remains to be evaluated. Finally, normalization appeared as a major bottleneck, especially because it relied mainly on context-free methods that were not supervised on the specific writing style of AP-HP discharge summaries. The next chapter therefore examines this problem more directly, by moving from surface-form matching toward context-aware biomedical entity linking.

CHAPTER 3

Synthetic Contextualized Augmentation for Biomedical Entity Linking

This chapter presents SynCABEL, a synthetic data augmentation framework for biomedical entity linking. The objective is to reduce the dependence of context-aware normalization models on costly expert annotations by generating contextualized training examples from biomedical knowledge bases.

1 Introduction

1.1 Context

The previous chapter showed that information extracted from discharge summaries can enrich structured electronic health record (EHR) data for patient phenotyping. However, it also highlighted an important limitation of information extraction pipelines: detecting a mention in text is not sufficient. To make extracted information reusable at scale, they must also be mapped to standardized concepts. This step, referred to as biomedical entity linking (BEL) or entity normalization, is essential for making extracted information queryable and reusable alongside structured data.

In Chapter 2, normalization was addressed with two types of methods: fuzzy matching methods (Levenshtein, 1966; Winkler, 1990), which compare character strings, and neural bi-encoder methods (Yuan et al., 2022c; Liu et al., 2021b), which compare dense representations of the detected mention and candidate concept names. These approaches differ technically, but they share an important limitation: the normalization decision is based only on the detected mention, without using the surrounding context. This can be problematic when the same surface form is ambiguous. For example, as illustrated in Figure 3.1, the abbreviation *AS* may refer to aortic valve stenosis in one sentence and to ankylosing spondylitis in another. In such cases, the correct concept cannot be selected reliably from the mention alone. It depends on the clinical context in which the mention appears.

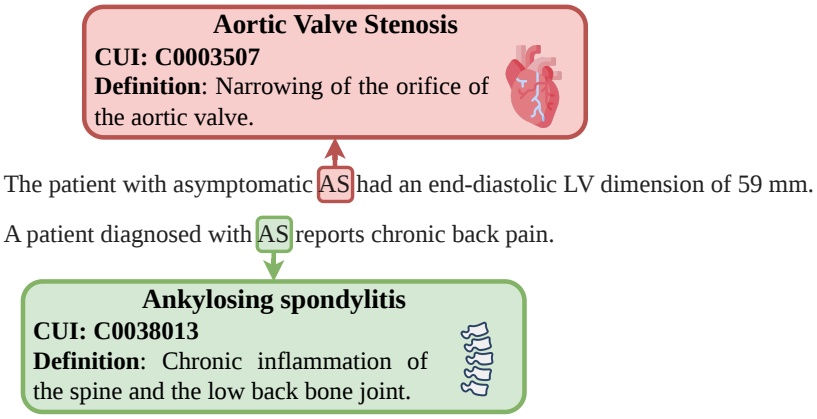


Figure 3.1: Example of contextual ambiguity in biomedical entity linking. The same abbreviation, *AS*, can refer to different biomedical concepts depending on the surrounding context, such as aortic valve stenosis or ankylosing spondylitis.

This motivates the use of context-aware BEL models, which link a mention by taking into account both the mention itself and the words that surround it. These models are particularly useful for ambiguous abbreviations and context-dependent expressions. However, they require annotated examples in which mentions are observed in context and linked to the correct concept. Such annotations are expensive to produce, because they require expert knowledge and careful selection among many possible concepts in large knowledge bases (KBs), such as the Unified Medical Language System (UMLS) (Bodenreider, 2004) or Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) (Donnelly, 2006) which contain several million concepts. As a result, human-annotated corpora usually cover only a small fraction of the target KB. A supervised model may therefore learn to disambiguate concepts that are present in the training set, while still failing on concepts that were never annotated. This chapter addresses this limitation by introducing SynCABEL, a synthetic contextualized augmentation strategy designed to increase the coverage of context-aware BEL training while reducing the dependence on manually annotated data.

1.2 Objectives and Contributions

The objective of this chapter is to evaluate whether synthetic contextualized data can reduce the annotation bottleneck in BEL. More specifically, we investigate whether examples generated from a KB can improve context-aware BEL models, especially for concepts that are absent from the human-annotated training set.

The intuition behind SynCABEL is summarized in Figure 3.2. Human-annotated data provide realistic contexts but cover only a limited number of concepts. KBs provide broad concept coverage, but they do not directly provide contextual examples. SynCABEL aims to combine both advantages by generating contextualized examples for KB concepts and using them to train context-aware BEL models.

The chapter makes three main contributions. First, we introduce SynCABEL, a framework that generates contextualized training examples from knowledge-base concepts and combines them with human annotations to train context-aware generative BEL models. Second, we use guided inference to restrict model predictions to valid con-

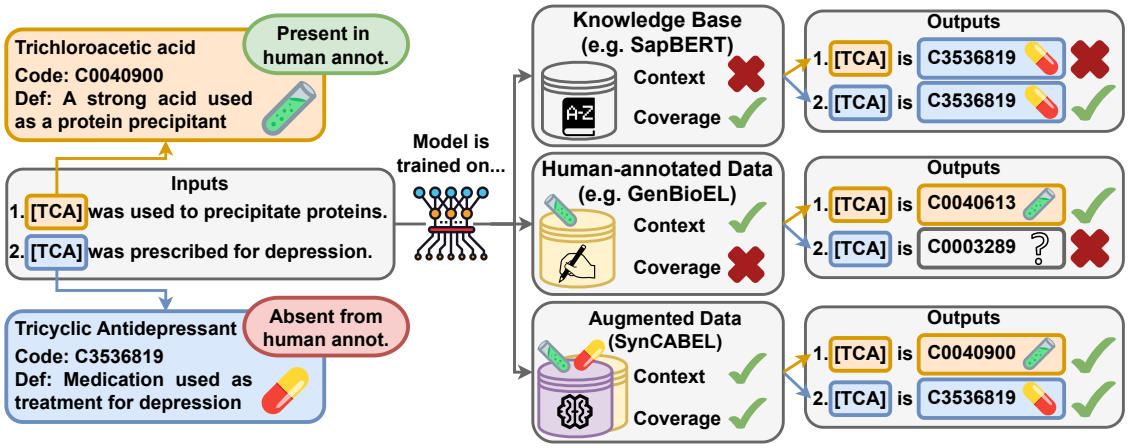


Figure 3.2: Illustration of the motivation behind SynCABEL. Human-annotated data provide contextual examples but limited concept coverage, while knowledge bases provide broad coverage but no direct contextual supervision. SynCABEL generates contextualized examples from knowledge-base concepts to increase the coverage of context-aware biomedical entity linking models.

cepts in the target KB. Third, we evaluate the approach on multilingual biomedical BEL benchmarks and complement exact code matching with an evaluation of clinical validity using a large language model (LLM)-as-a-judge.

1.3 Related Work

Chapter 1 introduced the main families of BEL methods, including rule-based systems, context-free bi-encoders, contextual bi-encoders, rerankers, generative models, and recent approaches based on large language models. To avoid repeating this general overview, this section focuses on the aspects most directly related to SynCABEL: context-aware BEL and strategies for creating additional training data.

1.3.1 Context-aware Biomedical Entity Linking

Context-aware BEL models were developed for cases where the mention surface form alone is not sufficient to select the correct concept. Instead of comparing only the detected mention with concept names or synonyms, these models also use the sentence or document context in which the mention appears. This is particularly useful for abbreviations and expressions whose meaning depends on the surrounding words.

A first family of context-aware models relies on contextual bi-encoders. In this setting, the mention is encoded together with its context, while each candidate concept is encoded from information available in the KB, such as its name, synonyms, or definition. The model then links the mention to the candidate concept whose representation is closest in the embedding space (Wu et al., 2020; Bhowmik et al., 2021). This formulation remains efficient because concept representations can be precomputed, while the mention representation can still depend on its context.

Several works have extended this idea by using richer training objectives. ArboEL, for example, does not only learn to compare each mention independently with candidate

concepts. It also models relations between mentions, so that mentions that are likely to refer to the same concept can support one another during training and prediction (Agarwal et al., 2022). This is useful because biomedical texts often contain several mentions of the same entity, sometimes with different surface forms. Instead of making each linking in isolation, ArboEL encourages a more coherent interpretation of related mentions.

Another family of context-aware methods formulates BEL as a generation task. Given a mention and its context, the model generates the corresponding concept name or identifier. This formulation was introduced in general entity linking by GENRE, where constrained decoding ensures that the generated output corresponds to a valid entity in the target KB (Cao et al., 2021a). In BEL, GenBioEL adapted this idea to biomedical KBs by using knowledge-base-guided pretraining and synonym-aware fine-tuning (Yuan et al., 2022b). Generative models are attractive because they provide a direct sequence-to-sequence formulation of BEL and can use guided inference to restrict predictions to valid concepts.

These approaches show that context is valuable for BEL. However, they also make the need for supervision more important. To learn how context determines the correct concept, the model requires examples where mentions are observed in context and linked to reference concepts. These annotations are expensive to produce and usually cover only a small fraction of the target KB. This motivates the use of data augmentation strategies.

1.3.2 Data Augmentation for Biomedical Entity Linking

Data augmentation aims to reduce the dependence on manually annotated mention–concept pairs. A common strategy is weak or distant supervision. In this setting, mentions are automatically identified in large text corpora by matching words or phrases to concept names and synonyms from a KB. This makes it possible to create large training datasets without manual annotation (Zhang et al., 2022; Vashishth et al., 2021). However, the resulting labels can be noisy. A matched string does not always refer to the expected concept, especially when the surface form is ambiguous or when the context differs from the target biomedical or clinical setting.

Another strategy is template-based synthetic data. Instead of searching for concept names in existing corpora, one can create artificial examples from the KB itself. For example, concept definitions or predefined sentence templates can be filled with concept names or synonyms (Yuan et al., 2022b). This gives more control over the target concept, because the label is known at generation time. However, template-based examples may lack the lexical and contextual diversity of natural biomedical text. They can help the model learn concept names and definitions, but they may not fully reflect the variety of contexts in which mentions appear in practice.

Large language models provide a more flexible way to generate synthetic examples. They can produce natural sentences, vary the mention form, and create more diverse contexts than fixed templates. Recent work has explored this direction for BEL (Chen et al., 2025a; Xin et al., 2025). LLMs have also been used in related parts of BEL pipelines, for example to rewrite complex mentions before linking or to select the best concept among retrieved candidates (Borchert et al., 2024; Xie et al., 2024; Lin et al., 2024). These approaches show that LLMs can support BEL in different ways, but using them directly at inference time can be costly and difficult to control at scale.

For data augmentation, an important limitation is that many LLM-based approaches

generate additional examples mainly for concepts already present in the human-annotated training data. This can improve contextual diversity for known concepts, but it does not fully address the lack of supervision for concepts that never appear in the annotated corpus. The central challenge is therefore not only to generate more examples, but to generate useful contextualized examples that increase the coverage of the target KB.

1.3.3 Positioning of SynCABEL

SynCABEL is positioned at the intersection of context-aware BEL and synthetic data augmentation. Like supervised context-aware models, it aims to use the surrounding text to disambiguate mentions. Unlike standard supervised approaches, it does not rely only on the concepts that happen to appear in a human-annotated corpus. Instead, it generates contextualized examples for concepts in the target KB, using information such as concept names, synonyms, semantic groups, semantic types, and definitions.

The objective is not to replace expert annotation. Human-annotated examples remain essential because they reflect the target corpus and provide high-quality supervision. SynCABEL instead aims to increase the coverage and diversity of the training data, so that context-aware BEL models can learn from both real annotated examples and synthetic examples generated from the KB. In this way, the approach addresses a central limitation of BEL: the need to use context while keeping sufficient coverage of a large concept space.

The next section presents the SynCABEL methodology in detail, including the formulation of generative BEL, the construction of synthetic contextualized examples, the combination of synthetic and human-annotated data, and the guided inference procedure used to restrict predictions to valid concepts.

2 Methods

2.1 Problem Statement

Let $\mathcal{E}$ represent the set of candidate concepts from a given KB. Each concept $e \in \mathcal{E}$ is associated with a semantic group g_e (e.g., *Disorders*) and multiple synonyms $S_e = \{s_e^1, \ldots, s_e^{n_e}\}$. Given a mention m with left context c_l, right context c_r and semantic group g_e, our objective is to predict the correct concept e_m that the mention m refers to.

2.2 Generative BEL Training Objective

We adopt the autoregressive formulation from Cao et al. (2021a). The input sequence x is defined as:

$$x = c_l \; [\; m \;] \; \{\; g_e \;\} \; c_r \; [\; SEP \;] [\; m \;] \text{is}$$

where c_l, c_r denote context, m the target mention, and g_e its semantic group. The token $[SEP]$ and cue phrase "m is" prompt the model to generate output sequence y defined as:

$$y = e_m$$

where e_m is the concept corresponding to mention m. We maximize the conditional likelihood:

$$p_\theta(y|x) = \prod_{i=1}^{N_y} p_\theta(y_i|y_{<i}, x)$$

where N_y is the number of tokens in the output, and y_i is the i-th token.

2.3 Adaptive Concept Representation for Training

A key challenge during training is representing the concept e_m in natural language for the target output. We employ an adaptive method that selects the most appropriate synonym of a concept based on the input mention. As shown in Figure 3.3, the method proceeds in two steps. First, we preprocess each concept's list of synonyms by removing those that are ambiguous with other concepts in the same semantic group. For example, the term "discharge" is removed from the *Disorder* group because it can refer to multiple clinical symptoms. Then, for each remaining unambiguous synonym s_e, we compute the cosine similarity between the embeddings of the mention m and s_e. The concept representation e_m is defined as the synonym with the highest similarity score.

$$e_m = \arg\max_{s_e \in \mathcal{S}_e} \frac{\mathbf{v}_m \cdot \mathbf{v}_{s_e}}{\|\mathbf{v}_m\| \|\mathbf{v}_{s_e}\|}$$

where $\mathbf{v}_m$ and $\mathbf{v}_{s_e}$ are the embedding vectors of the mention m and synonym s_e, respectively. Importantly, the embeddings can be derived from any model, in our experiments, we compare different embedding models to assess their impact on linking performance.

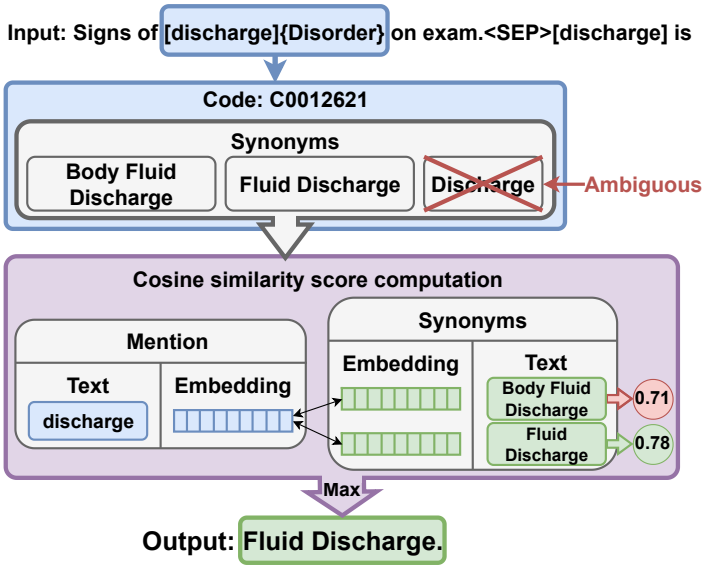


Figure 3.3: Example of the disambiguation process used to select a natural representation e_m = "Fluid Discharge" for a concept e = "C0012621" and a mention m = "discharge".

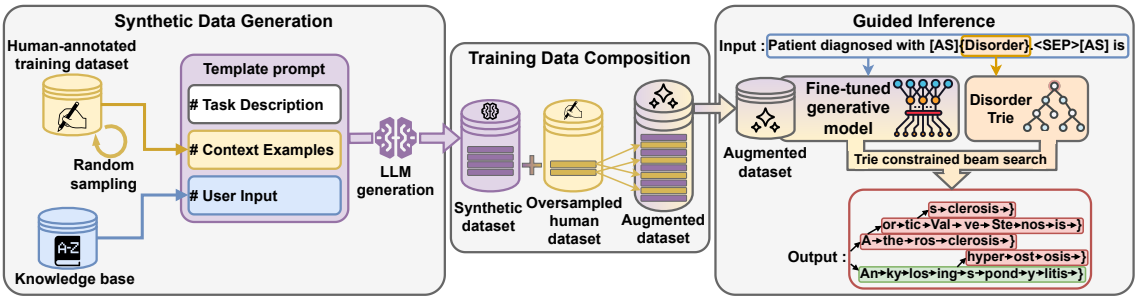


Figure 3.4: Overview of SynCABEL. LLM: large language model

2.4 Synthetic Data Generation

To improve concept coverage while maintaining contextual realism, we generate synthetic training data using a structured prompt-based approach. For a given concept e, our method generates contextual sentences that capture diverse, natural ways of expressing e. As illustrated in Figure 3.4, this is achieved by constructing a structured prompt composed of three key components: (i) the task description that explicitly structures the generation process into two sequential tasks: first, selecting and generating mentions of the concept. Second, generating contextualized sentences for each generated mention, aligned with the style of the provided human-annotated dataset examples. (ii) Random contextual examples from a human-annotated training set (iii) Title, semantic group, semantic type, definitions and synonyms of the target concept e from the KB, and this structured prompt is provided to an LLM, which produces the final contextualized sentences. This structured prompt is formulated in the same language as the target dataset.

2.5 Training Data Composition

The synthetic examples generated by SynCABEL are combined with human-annotated training data to fine-tune the BEL model (Figure 3.4). To preserve the characteristics of clinical reports while benefiting from the increased coverage of synthetic data, we jointly train on both sources and upsample human-annotated examples such that they constitute half of the training instances.

2.6 Guided Inference

Greedy decoding often yields invalid outputs absent from the KB. To address this, we adopt the guided inference mechanism from Cao et al. (2021a), illustrated in Figure 3.4. First, the KB vocabulary is filtered by the target entity's semantic group (e.g., *Disorders*) and structured as a trie, where each path maps a synonym to a unique concept. Then, the model dynamically restricts token selection at each step to valid branches. For instance, after generating the prefix "A", the trie permits only valid continuations like "the" (for *Atherosclerosis*) or "ortic" (for *Aortic*).

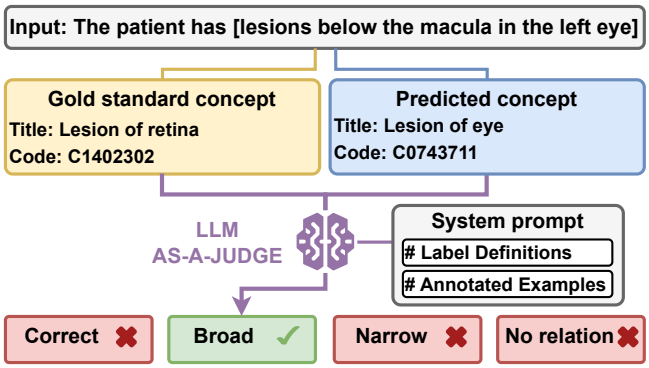


Figure 3.5: LLM-as-a-judge evaluation of predicted vs. gold concepts using four clinical-relation classes: Correct, Broad, Narrow, and No relation.

2.7 LLM-as-a-judge Evaluation

Standard entity linking evaluation relies on exact code matching and therefore may fail to capture the true clinical relationship between predicted and gold concepts. In practice, a prediction can be clinically correct despite a code mismatch due to ontology redundancy (e.g., multiple codes for the same concept) or annotation noise, where the prediction better reflects the mention than the human label. To capture these nuances, we introduce a four-class scheme (Figure 3.5): **(i) Correct**, clinically appropriate even without exact code match; **(ii) Broad**, relevant but overly general; **(iii) Narrow**, relevant but overly specific; and **(iv) No relation**, clinically unrelated. Labels are assigned automatically by an LLM using precise class definitions and five clinician-annotated examples.

3 Experiments

3.1 Datasets

3.1.1 Human-Annotated Datasets

We rely on three human-annotated corpora for our experiments:

- **MedMentions-ST21pv (MM-ST21pv)** (Mohan et al., 2019), a curated subset of the MedMentions corpus, consisting of 4,392 English PubMed abstracts. MedMentions is currently the largest publicly available human-annotated dataset for BEL.
- **QUAERO French Medical Corpus (QUAERO)** (Névéol et al., 2014), the largest French dataset for BEL, composed of two subsets: QUAERO-MEDLINE, derived from MEDLINE titles, and QUAERO-EMEA, extracted from documents published by the European Medicines Agency.
- **Spanish Clinical Case Corpus (SPACCC)** (Miranda-Escalada et al., 2022; Lima-López et al., 2023a; Lima-López et al., 2023b), a manually annotated collection of clinical case reports derived from open-access Spanish medical publications. The corpus contains 1,000 clinical cases, with disease, procedure, and symptom mentions annotated using the SNOMED CT ontology.

	MM-ST21PV	QUAERO-MEDLINE	QUAERO-EMEA	SPACCC
# Documents	4,392	2,384	26	1,000
Average length	1,581	84	5,062	2,337
Language	English	French	French	Spanish
Clinical cases	✗	✗	✗	✓
# Mentions	203,282	9,021	7,127	27,103
# Concepts	25,419	4,281	1,226	9,159
# Sem. Groups	14	10	10	3
Coverage	1.07%	0.14%	0.04%	3.34%

Table 3.1: Summary statistics of the human-annotated evaluation datasets. Average length: average document length in characters; Clinical cases: whether the dataset consists of clinical case reports; # Sem. Groups: number of UMLS semantic groups represented in the dataset; Coverage: proportion of concepts from the target terminology represented in the dataset. MM-ST21pv: MedMentions-ST21pv; SPACCC: Spanish Clinical Case Corpus.

	MM-ST21pv	QUAERO	SPACCC
Version	2017AA	2014AA	2021
# Concepts	2,374,392	2,960,365	276,670
# Synonyms	4,905,087	6,970,593	546,814
# Semantic Groups	14	10	3
# Synonyms per Code	2.1	2.4	1.5

Table 3.2: Summary statistics of the knowledge base subsets used for candidate retrieval across the evaluated datasets.

MM-ST21pv and QUAERO were selected because they are the largest human-annotated BEL datasets available in their respective languages and because they cover nearly all UMLS semantic groups. SPACCC was chosen as the largest Spanish BEL dataset and, more importantly, because it consists of clinical reports, making it particularly valuable for assessing the potential applicability of our approach in clinical settings. Table 3.1 summarizes the main characteristics of these datasets.

3.1.2 Knowledge bases

For each corpus, we selected a KB consistent with its original annotation setup. For MM-ST21pv, we use UMLS 2017AA, narrowed down to the ST21-pv subset which covers 21 semantic types with synonyms sourced from 18 vocabularies. QUAERO is paired with UMLS 2014AA, from which we retained only the 10 semantic groups relevant to the corpus. For the SPACCC dataset, candidate concepts were drawn from gazetteers built from the appropriate branches of the Spanish edition of SNOMED-CT (July 31, 2021 release), since SPACCC comprises three subsets (diseases, symptoms, procedures), these categories served as semantic groups. The statistics of these KB subsets are summarized in Table 3.2.

Dataset	SynthMM	SynthQUAERO	SynthSPACCC
# Mentions	460,878	396,914	1,813,463
# Concepts	153,374	132,089	276,649

Table 3.3: Summary of synthetic datasets.

Dataset	Concept	Semantic group	Generated sentence
SynthMM	C0013364 Dysautonomia, Familial	Disorders	"Genetic analysis revealed a significant association between mutations in the IKBKAP gene and the development of [Riley-Day Syndrome], a rare autosomal recessive disorder characterized by congenital sensory and autonomic dysfunction."
Synth QUAERO	C0033575 Prostatic Diseases	Disorders	"Les [Troubles prostatiques] peuvent avoir un impact significatif sur la qualité de vie, notamment en ce qui concerne la fonction sexuelle et la santé urinaire."
Synth SPACCC	C0203240 Radiography of wrist	Procedures	"En la sala de urgencias, se realiza una [placa radiográfica de muñeca] para diagnosticar una posible fractura por estrés."

Table 3.4: Examples of synthetic contextualized sentences generated by SynCABEL for the three target datasets. The concept column reports the concept identifier and title, and brackets indicate the generated mention used for entity linking.

3.1.3 Synthetic Datasets

We constructed three synthetic datasets (SynthMM, SynthQUAERO, and SynthSPACCC) to complement the human-annotated corpora. We employed Llama-3-70B (Grattafiori et al., 2024) to generate three training examples per concept, using prompts containing five random annotated examples from the target training set. For SynthMM and SynthQUAERO, generation was restricted to UMLS concepts with available definitions, covering 6.5% and 4.5% of the KB, respectively, to maintain computational feasibility. In contrast, for SynthSPACCC, examples were generated for all candidate concepts because the target KB was smaller. This illustrates a practical trade-off between synthetic coverage and generation cost when scaling to very large biomedical KBs.

Table 3.3 summarizes the main statistics of the generated datasets. Examples of generated sentences are provided in Table 3.4. Brackets indicate the generated mention used as the target for entity linking. The full datasets are available on Hugging Face, together with the prompts used for data generation. Prompts are also provided in Section B.2 of Appendix B.

3.2 Implementation Details

All experiments were conducted on a single NVIDIA H100 GPU with 80 GB of memory. The codebase was implemented in Python using PyTorch and the Hugging Face Transformers library. Generative models were fine-tuned using the `trl SFTTrainer` module. The main training and inference hyperparameters are reported in Tables 3.5 and 3.6. For reproducibility, the full code and experimental configuration are available on GitHub.

Hyperparameter	**Value**
Learning rate	3×10^{-5}
Learning rate scheduler	Linear
Warmup ratio	0.03
Epochs	50
Precision	`BF16`
Attention implementation	Flash Attention 2
Gradient accumulation steps	1

Table 3.5: Supervised fine-tuning hyperparameters used across the evaluated generative models.

Hyperparameter	**Value**
Decoding strategy	Beam search
Number of beams	5
Batch size	16
Precision	`BF16`

Table 3.6: Inference hyperparameters used during concept generation.

3.3 Results

3.3.1 Benchmark Performance

We evaluated two families of BEL systems. The first family is context-free BEL, where the mention is linked mainly from its surface form and the labels or synonyms available in the KB. This includes the rule-based **SciSpacy** (Neumann et al., 2019) and the context-free bi-encoders **SapBERT** (Liu et al., 2021a), **SapBERT-all** (Liu et al., 2021b), and **CODER-all** (Yuan et al., 2022c). The second family is supervised context-aware BEL, where models are trained on human-annotated data and use the textual context surrounding the mention. This includes the contextual bi-encoder **ArboEL** (Agarwal et al., 2022) and generative models based on **mBART-large** (Tang et al., 2020), **Llama-3-1B**, and **Llama-3-8B** (Grattafiori et al., 2024). In the generative setting, SynCABEL can be applied to different backbone models. We considered one encoder–decoder model and two decoder-only models. All generative models use the constrained decoding strategy introduced by GENRE (Cao et al., 2021a), which restricts generation to valid concepts from the target KB. The encoder–decoder backbone was selected through a preliminary

Model	MM-ST21PV (english)	QUAERO-MEDLINE (french)	QUAERO-EMEA (french)	SPACCC (spanish)	Average
CONTEXT-FREE BEL					
SciSpacy (Neumann et al., 2019)	53.8±1.0	40.5±1.9	37.1±4.3	13.2±0.9	36.2
SapBERT (Liu et al., 2021a)	65.6±1.0	63.2±1.8	59.7±3.8	33.9±1.4	55.6
CODER-all (Yuan et al., 2022c)	62.9±1.1	72.5±1.8	66.9±4.0	43.7±1.4	61.5
SapBERT-all (Liu et al., 2021b)	64.6±1.1	74.7±1.8	67.9±3.9	47.9±1.5	63.8
BERGAMOT (Sakhovskiy et al., 2024)	60.9±1.1	71.4±1.8	63.8±4.9	46.0±1.5	60.5
CONTEXT-AWARE BEL					
ArboEL (Agarwal et al., 2022)	**76.9**±0.9	71.0±1.9	63.0±3.9	49.0±1.7	65.0
mBART-large (Tang et al., 2020)	69.6±1.0	72.8±2.1	69.3±5.4	61.8±2.7	68.4
+ SynCABEL (Our method)	71.5±1.0	77.1±1.9	75.3±4.3	64.0±2.3	72.0
Llama-1B (Grattafiori et al., 2024)	73.1±1.0	76.2±1.4	75.1±3.6	63.8±1.6	72.1
+ SynCABEL (Our method)	74.2±0.9	77.0±1.6	75.8±3.0	65.0±1.6	73.0
Llama-8B (Grattafiori et al., 2024)	75.0±0.9	77.6±1.5	73.8±3.9	65.1±1.5	72.9
+ SynCABEL (Our method)	75.9±0.8	**79.7**±1.4	**79.0**±3.9	**67.0**±1.3	**75.4**

Table 3.7: Entity linking performance (Recall@1) on biomedical benchmarks. The highlighted row shows our contribution, where the base model is trained on an augmented dataset. The best results are shown in **bold**, the second-best results are underlined, and the "Average" column reports the mean score across the four benchmarks. 95% confidence intervals are estimated using empirical bootstrap resampling at the document level.

comparison on MM-ST21pv and QUAERO, reported in Table S9 in Appendix B. This comparison showed that guided inference substantially improved encoder–decoder models, and that mBART-large achieved the strongest overall performance among the tested backbones. We therefore retained mBART-large as the encoder–decoder backbone. For decoder-only models, we evaluated Llama-3-1B and Llama-3-8B directly.

In the main benchmark, we compared two configurations for each generative backbone: supervised fine-tuning on the human-annotated data with guided inference, and the same setting augmented with SynCABEL synthetic examples. This design isolates the contribution of SynCABEL while keeping the inference procedure fixed. For fair comparison, all systems except SciSpacy were constrained to the same KBs, and for each mention, only concepts from the matching semantic group were considered as candidates. In practice, this meant that for a given input, each system was provided with the same candidate set. This constraint could not be applied to SciSpacy, because its rule-based approach uses the entire KB and cannot be modified.

Table 3.7 presents BEL performance (Recall@1) across four biomedical benchmarks: MM-ST21pv, QUAERO-MEDLINE, QUAERO-EMEA, and SPACCC. Models are grouped according to whether they use the mention context. SynCABEL with Llama-3-8B achieves the highest scores on all four benchmarks: 75.9 on MM-ST21pv, 79.7 on QUAERO-MEDLINE, 79.0 on QUAERO-EMEA, and 67.0 on SPACCC, with the best average score of 75.4. The mBART-large and Llama-3-1B variants of SynCABEL also show consistent improvements over their corresponding guided-inference baselines.

	MM-ST21PV	QUAERO-MEDLINE	QUAERO-EMEA
Title (Static)	61.7	53.7	50.6
CODER-all	68.4	72.5	66.0
TF-IDF	**69.6**	**72.8**	**69.3**

Table 3.8: Recall@1 scores for different concept representation methods. mBART-large was supervised only on the human-annotated dataset (i.e., without data augmentation), and inference was performed using the guided inference method. The highlighted row shows the chosen representation method.

3.3.2 Adaptive Concept Representation for Training

We evaluate our adaptive concept representation strategy against a static baseline based on concept preferred titles. We compare two adaptive representations: character-level 3-gram TF-IDF (Neumann et al., 2019) used in GenBioEL (Yuan et al., 2022b) and CODER-all embeddings (Yuan et al., 2022c). Experiments are conducted on UMLS-based datasets (MM-ST21pv, QUAERO-MEDLINE and QUAERO-EMEA) using mBART-large, trained on the human-annotated data and decoded with guided inference. As shown in Table 3.8, the TF-IDF-based representation consistently achieves the highest Recall@1 across all datasets, outperforming both the static baseline and CODER-all. We therefore adopt TF-IDF-based adaptive representations for all other experiments.

3.3.3 Training Data Composition

We evaluate three training data composition strategies when combining human-annotated and synthetic data: **(i) Synthetic Pretrain**, which uses a two-stage training procedure, first training on synthetic data and then fine-tuning on human data, without resampling; **(ii) Combined**, which performs a single-stage training on the union of human and synthetic datasets as-is, without resampling; **(iii) Interleaved**, which performs a single-stage training on merged data while oversampling human samples so that each training step alternates between a human and a synthetic instance. These experiments use Llama-3-8B with guided inference.

	MM-ST21pv	QUAERO-MEDLINE	QUAERO-EMEA	SPACCC
SPT	74.1	78.0	77.1	64.0
COMB	75.4	79.4	77.0	64.3
INT	**75.9**	**79.7**	**79.0**	**67.0**

Table 3.9: Recall@1 scores for different training data composition strategies. The highlighted row shows the chosen strategy.

Recall@1 results in Table 3.9 show that interleaved oversampling consistently outperforms other strategies across all datasets, highlighting the importance of preserving a strong human signal while leveraging synthetic diversity. We therefore adopt this strategy in other experiments.

3.3.4 LLM-as-a-Judge Evaluation

Exact code matching provides a strict evaluation of BEL systems, but it may underestimate clinical validity. In some cases, a prediction can differ from the reference code while still referring to the same clinical concept, or to a clinically acceptable broader or narrower concept. To quantify this effect, we first manually analyzed a subset of code mismatches on SPACCC, the only dataset containing full clinical reports. A clinician annotated 150 mismatches produced by our best-performing model, Llama-3-8B, with 50 examples sampled from each of three semantic groups: disorders, symptoms, and procedures. Each mismatch was assigned one of four labels: *Correct*, when the predicted concept was clinically equivalent to the reference despite the code mismatch; *Broad*, when the predicted concept was more general; *Narrow*, when it was more specific; and *No relation*, when it was not clinically related.

The distribution of clinician labels is shown in Table 3.10. Overall, 26.7% of the mismatches were judged as *Correct*. Beyond these correct predictions, many mismatches were still clinically related to the reference concept: 28.7% were judged as broader and 21.3% as narrower. Only 23.3% of the sampled mismatches were labeled as *No relation*. This confirms that exact code matching alone substantially underestimates the clinical relevance of model predictions: 76.7% of the reviewed mismatches retained at least some semantic relation to the reference concept.

Semantic group	**Correct**	**Broad**	**Narrow**	**No relation**
DISO (N = 50)	15 (30.0%)	17 (34.0%)	8 (16.0%)	10 (20.0%)
PROC (N = 50)	16 (32.0%)	13 (26.0%)	11 (22.0%)	10 (20.0%)
SYMP (N = 50)	9 (18.0%)	13 (26.0%)	13 (26.0%)	15 (30.0%)
Overall (N = 150)	**40 (26.7%)**	**43 (28.7%)**	**32 (21.3%)**	**35 (23.3%)**

Table 3.10: Distribution of clinician labels among code mismatches in the manually annotated subset of SPACCC. Percentages are computed within each semantic group, and overall percentages are computed over the full annotated subset.

These results motivated the use of an LLM-as-a-judge evaluation to complement exact code matching at a larger scale. We employed GPT-5.2[1] as an automated judge to assess the clinical relation between predicted and reference concepts on SPACCC. Before using the LLM-as-a-judge, we also explored a KB-based strategy to explicitly detect broader and narrower relations between predicted and reference concepts. However, this strategy did not yield reliable results, likely because biomedical KB hierarchies are complex and brittle for this purpose.

We evaluated the reliability of the LLM judge against the clinician annotations on the 150 manually reviewed mismatches. GPT-5.2 achieved 66.2% agreement with the clinician labels, with 79.0% precision for the *Correct* label. Its errors were mostly conservative: only 12.4% of judge predictions overstated the human label, for example by predicting *Narrow* instead of *No relation*. This tendency to under-predict relevance suggests that the judge provides a conservative estimate of clinical validity.

Table 3.11 compares standard exact matching with the LLM-as-a-judge evaluation. SynCABEL improves exact matching by 1.9 points, from 65.1% to 67.0%. The gain is

[1]Developed by OpenAI and accessed via their API.

larger for clinically valid predictions, with the *Correct* label increasing by 3.7 points, from 68.9% to 72.6%. At the same time, the proportion of predictions judged as having no clinical relation decreases by 3.3 points, from 13.8% to 10.5%. These results suggest that SynCABEL helps the model capture the correct clinical semantics even when the predicted concept does not exactly match the reference code.

Label	**w/o SynCABEL**	**w/ SynCABEL**	**Δ score**
Exact matching	65.1% [63.7–66.7]	67.0% [65.5–68.6]	+1.9
Correct	**68.9% [67.4–70.4]**	**72.6% [71.2–74.1]**	**+3.7**
Broad	13.0% [12.1–13.8]	11.0% [10.3–11.8]	-2.0
Narrow	4.3% [3.8–4.7]	5.9% [5.3–6.5]	+1.6
No relation	13.8% [11.1–16.7]	10.5% [7.6–13.2]	-3.3

Table 3.11: Llama-3-8B performance on SPACCC with and without SynCABEL augmentation. Results are reported as percentages with 95% confidence intervals estimated via document-level empirical bootstrap, for exact code matching and each LLM-as-a-judge label. The Δ score column reports the absolute difference between the two settings in percentage points.

4 Discussion

4.1 Annotation Scarcity

To assess whether SynCABEL helps beyond concepts already observed in the human-annotated training data, we stratified performance according to whether each target concept was seen or unseen in the manual annotations. Seen concepts are those that appear at least once in the human-annotated training set, whereas unseen concepts are absent from these annotations. Importantly, some of these unseen concepts may still be introduced during training through SynCABEL-generated examples. This distinction therefore evaluates whether synthetic augmentation can improve coverage beyond the concepts manually annotated in the training corpus.

As shown in Table 3.12, SynCABEL consistently improves performance on unseen concepts across all benchmarks, with gains ranging from +3.7 points on MM-ST21pv to +12.2 points on QUAERO-EMEA. At the same time, performance on seen concepts remains stable, with only very small variations. This suggests that synthetic augmentation mainly improves generalization to less frequent concepts, without degrading the model's ability to recognize concepts already covered by manual annotations.

However, the gap between seen and unseen concepts remains large. For example, on SPACCC, performance reaches 83.0 on seen concepts but only 30.2 on unseen concepts after augmentation. This suggests that SynCABEL improves generalization to concepts that are missing from the human-annotated training data, but does not fully replace direct supervision. The results alone do not identify the exact cause of this remaining gap. One possible explanation is that human-annotated examples are naturally closer to the evaluation documents, while synthetic examples may use different surface forms or place the concept in different contexts.

The effect of SynCABEL also depends on the coverage of the synthetic generation

Subset	w/o SynCABEL	w/ SynCABEL	Δ score
MEDMENTIONS-ST21PV			
Overall	75.0 [74.1–75.9]	75.9 [75.0–76.7]	+0.9
Seen Concepts	82.2 [81.3–82.9]	82.3 [81.5–83.0]	+0.1
Unseen Concepts	47.3 [45.1–49.6]	51.0 [48.8–53.2]	**+3.7**
QUAERO-MEDLINE			
Overall	77.6 [76.1–79.2]	79.7 [78.0–81.2]	+2.1
Seen Concepts	90.0 [88.6–91.4]	89.7 [88.3–91.0]	-0.3
Unseen Concepts	59.5 [56.7–62.4]	65.1 [62.1–68.1]	**+5.6**
QUAERO-EMEA			
Overall	73.8 [69.8–77.7]	79.0 [75.7–82.2]	+5.2
Seen Concepts	92.1 [89.8–94.0]	92.0 [89.6–94.3]	-0.1
Unseen Concepts	50.2 [45.1–55.2]	62.4 [56.4–68.7]	**+12.2**
SPACCC			
Overall	65.1 [63.7–66.7]	67.0 [65.5–68.6]	+1.9
Seen Concepts	83.4 [82.2–84.6]	83.0 [81.8–84.2]	-0.4
Unseen Concepts	23.2 [21.4–24.9]	30.2 [28.1–32.4]	**+7.0**

Table 3.12: Exact-match Recall@1 performance on four biomedical benchmarks comparing two variants of Llama-3-8B: one trained only on the corresponding human-annotated dataset with guided inference, and one further enhanced with SynCABEL. Results are reported overall and for subsets defined by seen/unseen concepts. The Δ score column reports the absolute difference between the two settings, computed as w/ SynCABEL minus w/o SynCABEL on the reported score scale. 95% confidence intervals were estimated via the empirical bootstrap method at the document level.

process. For the UMLS-based datasets, MM-ST21pv and QUAERO, generation was restricted to concepts with available definitions. As a result, some concepts that are unseen in the human-annotated training data are also absent from the synthetic data. This limits the possible gain and suggests a natural direction for improvement: extending generation to concepts without definitions, for example by relying more on hierarchical information.

4.2 Reducing Annotation Effort

We then evaluated whether SynCABEL can reduce the amount of human annotation needed to reach strong performance. To do so, we varied the fraction of human-annotated training documents from 0% to 100%. For each fraction, subsets were created by randomly sampling documents, and we compared models trained only on human annotations with models augmented using SynCABEL.

Figure 3.6 shows that SynCABEL can substantially reduce annotation needs. The SynCABEL-augmented model reaches the full-data baseline using only 60% of the human annotations on MM-ST21pv and SPACCC, and 40% on QUAERO. The benefit is largest when few annotated documents are available, and then decreases as more hu-

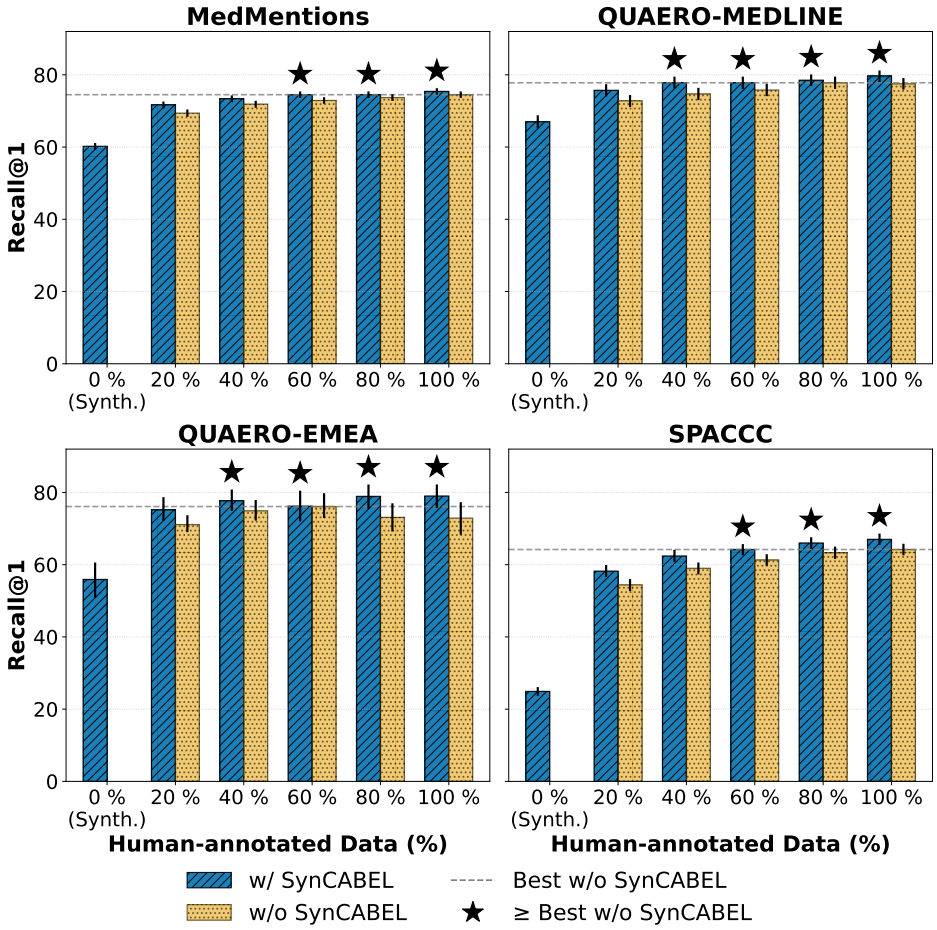


Figure 3.6: Data-efficiency analysis. Exact-match Recall@1 performance as a function of human training data availability. The dashed line marks the performance of the standard baseline trained on 100% of the human-annotated data. Stars ($\star$) indicate where the SynCABEL-augmented model reaches or exceeds this baseline. The 0% setting corresponds to training on synthetic data only. Error bars denote 95% bootstrap confidence intervals.

man data is added. This trend is clear on MM-ST21pv and SPACCC, while it is less visible on QUAERO because the confidence intervals are wider.

At the same time, models trained only on synthetic data remain below the fully supervised baselines. This is an important result: synthetic data is useful for reducing annotation effort, but it does not replace human annotations. Instead, SynCABEL is best understood as a way to make limited annotated datasets more effective.

4.3 Practical Considerations: Precision and Inference Footprint

The previous analyses show that SynCABEL improves both exact-match performance and clinically valid predictions, while also reducing the need for human annotations. Beyond these evaluation results, it is important to consider how such models could be used in practice as part of an information extraction pipeline. We therefore examine two additional practical aspects: whether predictions can be filtered according to their

Model	Model size (GB)	Candidate size (GB)	Speed (mention/s)
SapBERT	2.1	20.1	**575.5**
ArboEL	**1.2**	7.1	38.9
mBART	2.3	**5.4**	51.0
Llama-3-1B	2.4	**5.4**	69.6
Llama-3-8B	28.6	**5.4**	38.2

Table 3.13: Inference speed and memory footprint. Model size refers to the memory required by model parameters, candidate size refers to the memory required by the candidate representation, and speed is reported in mentions per second.

confidence, and whether inference remains feasible at scale.

4.3.1 Confidence-based Filtering

In biomedical applications, the desired precision–recall balance depends on the use case. Exploratory corpus analysis may tolerate some errors, whereas cohort construction or extraction of research variables often requires higher precision, since false positives can directly affect the resulting population or analysis.

The generative formulation used in this chapter provides a simple way to control this trade-off. Each prediction is associated with a confidence score, computed as the mean token probability of the generated concept name. This score can be used to filter predictions: a high threshold keeps only the most confident predictions, increasing precision at the cost of recall.

On SPACCC, a confidence threshold of 0.9 yielded a precision of 80.8% [79.6–81.9], a recall of 61.9% [60.2–63.5], and an F1-score of 70.1% [68.5–71.6]. This threshold should not be interpreted as universal, as the optimal value depends on the dataset, terminology, and downstream task. However, it shows that confidence scores can make the model more controllable, allowing users to select a precision-oriented operating point when reliability is more important than exhaustiveness.

4.3.2 Inference Footprint

Scalability is another important consideration. Biomedical KBs contain millions of concepts and synonyms, and a clinical data warehouse may contain millions of mentions to normalize. A BEL model should therefore be considered not only in terms of accuracy, but also in terms of memory use and inference speed.

Table 3.13 illustrates the different trade-offs between model families. SapBERT is by far the fastest model, but it requires a large dense candidate index and does not use mention context. ArboEL uses a smaller model and candidate representation, but inference is slower because of its reranking step. Generative models use a compact candidate trie, which is smaller than the dense index used by SapBERT. This is one advantage of constrained generation: the candidate representation remains compact while still restricting outputs to valid concepts.

Among generative models, Llama-3-1B offers a strong practical compromise, with a smaller footprint and faster inference than Llama-3-8B. In contrast, Llama-3-8B achieves

Dataset	Pipeline setting	Precision	Recall	F1-score
MM-ST21pv	*NER only* (He et al., 2022)	69.0 ± 0.9	67.9 ± 0.9	68.5 ± 0.8
	NER + CODER-all	60.5 ± 1.0	43.7 ± 0.9	50.7 ± 0.9
	NER + SapBERT-all	58.9 ± 1.1	42.0 ± 1.0	49.0 ± 1.0
	NER + SynCABEL	$\mathbf{61.0 \pm 1.0}$	$\mathbf{50.7 \pm 1.0}$	$\mathbf{55.4 \pm 0.9}$
QUAERO-MEDLINE	*NER only* (Touchent et al., 2024)	72.1 ± 1.8	57.9 ± 1.7	64.2 ± 1.5
	NER + CODER-all	67.2 ± 2.2	42.0 ± 1.6	51.7 ± 1.8
	NER + SapBERT-all	63.8 ± 2.3	40.8 ± 1.8	49.8 ± 1.9
	NER + SynCABEL	$\mathbf{71.4 \pm 1.9}$	$\mathbf{46.6 \pm 1.7}$	$\mathbf{56.4 \pm 1.7}$
QUAERO-EMEA	*NER only* (Touchent et al., 2024)	79.9 ± 2.9	65.3 ± 5.5	71.9 ± 3.9
	NER + CODER-all	72.2 ± 4.6	44.9 ± 5.6	55.4 ± 5.2
	NER + SapBERT-all	60.6 ± 3.8	44.2 ± 5.0	51.1 ± 4.1
	NER + SynCABEL	$\mathbf{74.5 \pm 2.7}$	$\mathbf{56.4 \pm 5.9}$	$\mathbf{64.2 \pm 4.3}$
SPACCC	*NER only* (Carrino et al., 2022)	76.3 ± 1.4	57.6 ± 2.0	65.7 ± 1.6
	NER + CODER-all	59.2 ± 1.7	26.7 ± 1.4	36.8 ± 1.6
	NER + SapBERT-all	55.3 ± 1.8	28.3 ± 1.5	37.4 ± 1.6
	NER + SynCABEL	$\mathbf{70.5 \pm 1.9}$	$\mathbf{38.3 \pm 1.9}$	$\mathbf{49.7 \pm 1.9}$

Table 3.14: End-to-end named entity recognition (NER) and normalization performance. For each dataset, the *NER only* row reports mention detection performance before normalization. The following rows add a normalization model on top of the same NER predictions. A prediction is considered correct only when both the entity mention and the normalized concept are correctly identified. Nested entities with the same label were filtered before evaluation by retaining only the outermost span.

the best linking performance in our experiments, but requires substantially more GPU memory and has lower throughput. This illustrates a classical trade-off: larger models can improve accuracy, while smaller models may be preferable when resources or processing time are limited. Overall, there is no single best model for all settings. The choice depends on the downstream objective, available hardware, and the acceptable balance between accuracy, precision, recall, and speed.

4.4 End-to-End Impact of Context-Aware Normalization

The main experiments in this chapter evaluate BEL with gold entity mentions, in order to isolate the normalization step. To assess whether these results also transfer to a practical information extraction setting, we additionally evaluate an end-to-end NER and normalization pipeline, closer to Chapter 2. The NER component follows the architecture introduced in Chapter 2 and illustrated in Figure 2.4: a BERT-based encoder combined with CNN and CRF layers. We select an encoder adapted to each dataset: DeBERTa-v3 (He et al., 2022) for MM-ST21pv, CamemBERT-bio (Touchent et al., 2024) for QUAERO-MEDLINE and QUAERO-EMEA, and BSC-Bio-EHR-ES (Carrino et al., 2022) for SPACCC. Detailed NER results by semantic group are reported in Tables S5 and S6 in Appendix A and Table S10 and S11 in Appendix B.

The same NER predictions are used for all normalization methods, so that differences in performance only reflect the normalization step. In this end-to-end setting, the NER model predicts the broad semantic group of each mention, and each normalization model searches only the corresponding subset of the KB. Unlike many NER evaluations that simplify the task by removing nested entities, we kept all annotated entities in the evaluation. The only filtering step concerns mentions whose gold concept code is absent from the target KB, which makes normalization impossible. For SynCABEL, we use the Llama-3-8B backbone, corresponding to the best-performing configuration. A prediction is considered correct only when the mention is detected and linked to the correct concept. Since each normalization model provides a confidence score, we select a threshold on the validation set by maximizing end-to-end F1-score over 1,000 candidate thresholds sampled from the empirical score distribution. This threshold is then fixed and applied to the test set. The corresponding precision, recall, and F1-score curves are reported in Figure S1 in Appendix B.

Table 3.14 shows that SynCABEL improves the complete pipeline, not only the gold-mention setting. It obtains the best end-to-end F1-score on all four datasets, with gains over the best non-SynCABEL baseline of $+4.7$ points on MM-ST21pv, $+4.7$ on QUAERO-MEDLINE, $+8.8$ on QUAERO-EMEA, and $+12.3$ on SPACCC. These gains are substantial and consistent across datasets.

4.5 Limitations

Although SynCABEL improves the coverage and data efficiency of context-aware BEL, several limitations remain.

4.5.1 Evaluation on Public Benchmarks

A first limitation is that the end-to-end comparison was conducted on public benchmarks rather than on the French discharge summaries used in Chapter 2. The previous subsection already provides a controlled comparison across several datasets, and the results are consistent: SynCABEL improves the full NER and normalization pipeline on all benchmarks. However, these datasets remain different from the AP-HP clinical notes studied in Chapter 2. The most direct experiment would have been to replace the normalization component of that pipeline with SynCABEL and measure its impact on laboratory test and drug normalization in discharge summaries. This was not possible because the available AP-HP normalization annotations were limited in size and mainly designed for evaluation, rather than for training and validating a supervised normalization model. As a result, the improvement observed on public benchmarks still needs to be confirmed on French hospital discharge summaries.

4.5.2 The Gap Between Synthetic and Human Supervision

A second limitation is that synthetic supervision remains much less effective than human annotation when used alone. The data-efficiency experiments show that SynCABEL improves data efficiency when added to human annotations, but the synthetic-only setting remains substantially below the human-only setting.

4.5.2.1 Synthetic Concept Selection. One possible explanation is the concept selection process. In the present work, concept selection was mainly driven by practical constraints. For the UMLS-based datasets, MM-ST21pv and QUAERO, synthetic generation was restricted to concepts with available definitions, because generating examples for all UMLS concepts would be computationally expensive. In contrast, for SPACCC, synthetic examples were generated for all candidate concepts. These strategies are simple, but they may be suboptimal. In UMLS-based datasets, restricting generation to concepts with definitions leaves some concepts without synthetic examples. For example, in MM-ST21pv, 1,625 concepts remained unseen after augmentation, compared with 3,240 with human annotations only. SynCABEL therefore substantially reduces the coverage gap, but does not fully close it. Conversely, generating examples for all candidate concepts, as in SPACCC, may include many concepts that are unlikely to appear in the target benchmark and may introduce unnecessary noise.

To explore whether concept selection could be an important factor, we conducted a preliminary post-hoc analysis on SPACCC. Concept representations were obtained with CODER-all (Yuan et al., 2022c) by encoding all lexical variants of each concept, averaging them, and normalizing the resulting centroid. We compared the concepts appearing in the human-annotated dataset with random KB samples stratified to match the same semantic-group distribution. The comparison was performed using Maximum Mean Discrepancy (MMD), a kernel-based two-sample statistic for comparing empirical distributions (Gretton et al., 2012). The mean MMD^2 between human concepts and semantic-group-stratified random samples was 0.0118, whereas the mean MMD^2 between two independent stratified random samples was 0.0004. This difference was significant under an empirical randomization test ($p < 0.01$) and the full distribution of MMD^2 values is reported in Figure S2 in Appendix B. Thus, even after controlling for semantic group, human concepts remained substantially different from random KB concepts in the embedding space.

This preliminary result supports the intuition that the concepts appearing in human-annotated benchmarks occupy a specific region of the KB. This may partly come from the annotation process, since the terminology search tools used by annotators may make some concepts easier to find than others. Generating synthetic examples from concepts sampled across the KB without considering the target dataset may therefore be inefficient. A better strategy could be to sample concepts that are close to the human training distribution, while excluding those already seen during training. This could improve the coverage of likely held-out concepts and reduce the cost of generating examples that are unlikely to be useful. Whether it also improves performance remains to be evaluated.

4.5.2.2 Beyond Concept Selection. However, concept selection is only one possible explanation for the lower performance of synthetic-only training. Synthetic examples may also differ from human annotations in the surface forms selected by the LLM, or in the biomedical contexts in which concepts appear. Human annotations come from the same corpora as the evaluation examples and therefore naturally contain mentions, formulations, and contextual patterns that are closer to the test distribution. Future work should therefore also include an intrinsic evaluation of the generated examples, for instance by asking biomedical experts to assess their coherence, diversity, and realism. Such analyses could help determine whether the remaining gap between synthetic

and human supervision comes primarily from concept selection, generated mentions, generated contexts, or a combination of these factors.

4.5.3 Sentence-Level Context

A final limitation is related to the length of context used by SynCABEL. The method generates and uses sentence-level examples, which are well suited to train context-aware models but do not fully represent the structure of real biomedical documents. In biomedical texts, useful evidence may appear elsewhere in the document: the same concept may be introduced with a complete expression and later reused as an abbreviation, but other related mentions, clinical findings, procedures, or treatments can also provide clues for disambiguation. A model that links each mention independently may therefore miss signals that are only available at the document level. This may partly explain why ArboEL (Agarwal et al., 2022) performs better than SynCABEL on MM-ST21pv: its mention-level structure encourages more consistent predictions across related mentions, which is especially useful in scientific abstracts where concepts and related biomedical clues are often repeated across the document. This limitation motivates Chapter 4, which explores a complementary direction by using document-level context and previous predictions to improve consistency in BEL.

5 Conclusion

This chapter introduced SynCABEL, a synthetic contextualized augmentation framework for BEL. The starting point was the limitation identified in Chapter 2: normalization based only on mention surface forms is not sufficient when biomedical expressions are ambiguous and context-dependent.

SynCABEL uses information from KBs to generate contextualized training examples. These examples are combined with human annotations to train context-aware generative BEL models. Across multilingual benchmarks, the method improved Recall@1, especially for concepts that were not observed in the human-annotated training data. It also reduced the amount of human annotation needed to reach strong performance, while confirming that expert annotations remain essential.

The chapter also showed that exact code matching provides only a partial view of BEL quality. In the LLM-as-a-judge evaluation, many predictions that did not exactly match the annotated code were still clinically meaningful: some corresponded to another appropriate code, others to a broader or narrower related concept, while a remaining fraction was unrelated. This analysis showed that SynCABEL increased the proportion of clinically correct predictions more than what was captured by exact-match evaluation alone. Finally, the inference-footprint analysis highlighted a practical trade-off between performance and computational cost: larger generative models achieved stronger linking performance, whereas smaller or context-free models may be preferable when throughput, memory, or deployment constraints are more important.

Overall, this chapter addresses a key limitation of context-aware BEL: the limited coverage of human-annotated training data. SynCABEL shows that synthetic examples can expose models to additional concepts and improve generalization, especially for concepts absent from the human annotations. However, these examples remain centered on

single concepts and short local contexts. They therefore improve concept coverage, but do not model how evidence is distributed across a full biomedical document, through repeated mentions, abbreviations, definitions, or related entities. The next chapter studies this complementary direction with LongBEL, moving from sentence-level entity linking to document-level and document-consistent entity linking.

CHAPTER 4

Long-Context and Document-Consistent Biomedical Entity Linking

This chapter presents LongBEL, a document-level framework for biomedical entity linking. The objective is to move beyond sentence-level entity linking by using full-document context and previous predictions to improve consistency across a biomedical document.

1 Introduction

1.1 Context

The previous chapter introduced SynCABEL, a synthetic contextualized augmentation strategy designed to reduce the annotation bottleneck in biomedical entity linking (BEL). By generating contextualized examples from knowledge-base (KB) concepts, SynCABEL improved the coverage of context-aware models, especially for concepts absent from the human-annotated training data. However, it still links mentions mainly from short local contexts and does not explicitly exploit relations between mentions in the same document.

This remains a limitation of most current BEL systems. Mentions are often linked independently, using only the mention itself or the sentence in which it appears. However, useful information may be distributed across the document. A concept may first appear under its full name and later be referred to through an abbreviation, a synonym, or a shorter expression. Other parts of the document may also provide clues, such as related clinical findings, anatomical locations, or treatments.

Figure 4.1 illustrates this limitation. In the local-context setting, the abbreviation *PET* is difficult to interpret from the sentence alone and is incorrectly linked to *Petechial haemorrhage*. In contrast, the full document contains clinical cues such as pregnancy, severe hypertension, and proteinuria. Together with previous linking decisions, these cues support the correct interpretation of *PET* as *Pre-eclamptic toxemia*.

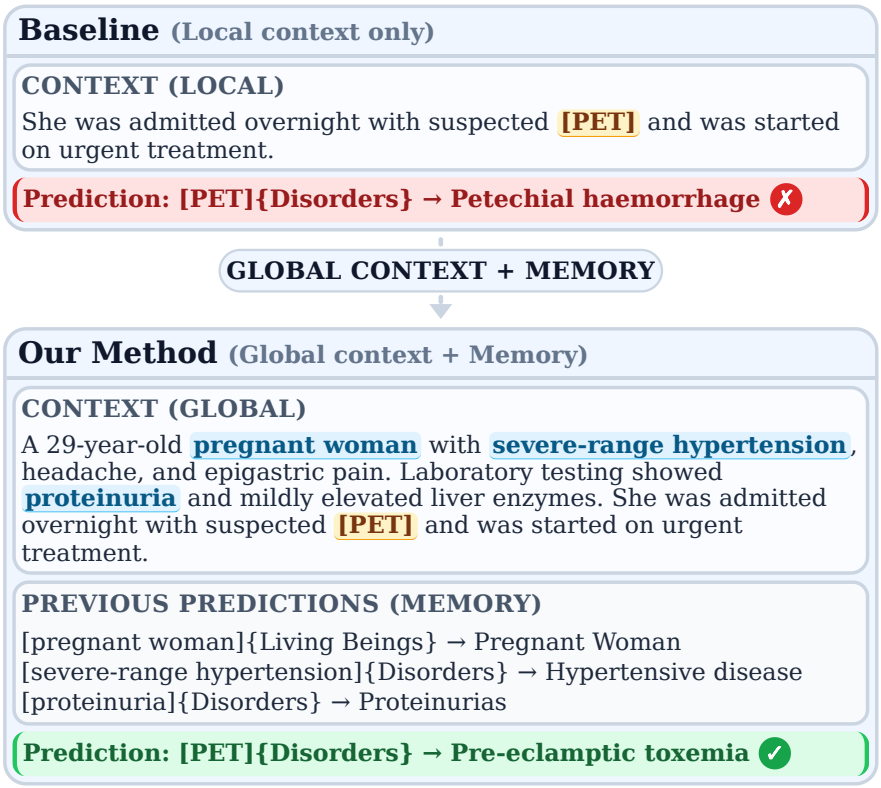


Figure 4.1: Illustrative example of document-level context and memory. **Top:** Restricted to local context, a standard model mislinks *[PET]* to *Petechial haemorrhage*. **Bottom:** By using global clinical context and previous predictions, LongBEL correctly links *[PET]* to *Pre-eclamptic toxemia*.

This issue is particularly important when BEL is used to transform text into structured variables for downstream analysis. The objective is not only to link isolated mentions correctly, but also to produce a coherent representation of the complete document. When the same concept appears under different surface forms, earlier mentions can provide useful evidence for later ones. This chapter therefore complements SynCABEL by studying how document-level context and previous predictions can be integrated into a generative collective BEL framework.

1.2 Objectives and Contributions

The objective of this chapter is to study whether a generative BEL model can benefit from document-level information. More specifically, we investigate whether full-document context and previous linking decisions can improve consistency across mentions in the same document.

To address this question, we introduce LongBEL, an autoregressive generative framework for collective BEL. LongBEL processes mentions sequentially and conditions each prediction on the full document and on a memory of previous predictions. This allows earlier linking decisions to support later mentions, especially when a concept recurs under an abbreviation, synonym, or shorter expression.

The contributions of this chapter are threefold. First, we implement collective BEL within a constrained generative framework, where document context and previous predictions directly influence concept generation. Second, we introduce a robust prediction memory built from out-of-fold model predictions rather than gold labels. This better matches inference conditions and reduces the risk of over-relying on an unrealistically perfect memory. Third, we evaluate LongBEL on three multilingual biomedical benchmarks in English, French, and Spanish. Our analyses show that its gains mainly concern recurring concepts and vary according to the characteristics of each dataset.

1.3 Related Work

Chapter 1 introduced the main families of BEL methods, including lexical systems, context-free bi-encoders, contextual bi-encoders, rerankers, generative models, and recent approaches based on large language models. Chapter 3 then focused on context-aware BEL and data augmentation. To avoid repeating these general overviews, this section focuses only on the aspects most directly related to LongBEL: long-context modeling and document-level consistency.

1.3.1 Document-Level Consistency and Long Context

Document-level consistency has long been studied through collective entity linking. The main idea is that mentions in the same document are not independent: they may belong to the same topic, refer to related entities, or repeat the same entity under different surface forms. Collective methods therefore use interactions among mentions and candidate entities rather than linking each mention separately.

Early approaches combined local predictions with coherence features derived from the relations between candidate entities (Ratinov et al., 2011). Later neural approaches mainly differ in how they combine local and document-level evidence. One approach uses a conditional random field to jointly score candidate entities from their learned representations (Ganea et al., 2017). Neural collective entity linking (NCEL) combines these signals through graph convolution (Cao et al., 2018), while recurrent random walk network learning (RRWNL) propagates evidence among related candidates through recurrent random-walk layers (Xue et al., 2019). In biomedical entity linking, neural methods also combine local and document-level information in different ways. A collective dual encoder processes the full document in one pass, allowing each mention representation to incorporate information from other mentions (Bhowmik et al., 2021). ArboEL instead explicitly models mention–mention and mention–entity relations through an arborescence objective (Agarwal et al., 2022). We therefore use ArboEL as the main collective BEL baseline in this chapter.

Most of these approaches are encoder-based and rely on joint representations, graphs, or structured inference. Recent open-weight language models have made another approach increasingly practical by supporting much longer inputs (Grattafiori et al., 2024). Generative BEL models can therefore use document-level context instead of being restricted to the sentence containing the target mention. Memory mechanisms have also been used to track entities across long narratives (Zhu et al., 2025), but they have not been directly studied for constrained generative BEL.

1.3.2 Positioning of LongBEL

LongBEL does not introduce the general idea of collective entity linking. Instead, it implements it within an autoregressive generative BEL framework. Mentions are processed sequentially, and each prediction is conditioned on the full document and on a memory of previous linking decisions. Constrained decoding ensures that every generated output corresponds to a valid concept in the target KB.

Unlike graph-based or retrieval-based collective methods, LongBEL does not use a separate candidate graph or collective reranking step. Cross-mention information is included directly in the model input and reused during generation.

LongBEL therefore addresses a different limitation from SynCABEL. SynCABEL improves concept coverage through synthetic short-context examples, whereas LongBEL focuses on document-level consistency by exploiting full-document context, recurring concepts, and previous predictions.

2 Methods

LongBEL extends the generative BEL approach introduced in Chapter 3 from local mention-level prediction to sequential document-level prediction. It retains the same concept representation and guided constrained decoding, but adds full-document context and a memory of previous linking decisions. This section describes the document-level formulation, the construction of a robust training memory, and the inference procedure.

2.1 Document-Level Autoregressive Formulation

Let $\mathcal{E}$ be the set of candidate concepts in the target knowledge base, and let c be a document containing an ordered sequence of detected mentions $\mathcal{M} = (m_1, m_2, \ldots, m_n)$. The index $t \in \{1, \ldots, n\}$ denotes the current prediction step, corresponding to the t-th mention in the document. Each mention m_t is associated with a semantic group g_t and must be mapped to a target concept $e_t \in \mathcal{E}$.

LongBEL follows the generative formulation introduced in Chapter 3. Mentions are processed sequentially in their order of appearance, and for each mention m_t, the model generates a textual representation of the target concept e_t. The novelty of LongBEL is that, at each step t, the prediction is conditioned on the full document c and on a memory $\mathcal{H}_t$ containing the mentions processed earlier and their predicted concepts:

$$\mathcal{H}_t = \{(m_1, \hat{e}_1), \ldots, (m_{t-1}, \hat{e}_{t-1})\}.$$

The memory is empty at the first step and is updated after each prediction.

In the example shown in Figure 4.1, the mention *PET* belongs to the *Disorders* semantic group. Its local sentence provides little evidence for selecting *Pre-eclamptic toxemia*, whereas the full document and previous predictions provide relevant information about pregnancy, hypertension, and proteinuria.

2.2 Training

2.2.1 Example Construction

Each example is defined by an input sequence x_t and a target sequence y_t. The input x_t is constructed from three elements: the full document c, the memory $\mathcal{H}_t$, and the current mention m_t with its semantic group g_t. This allows the model to use both the global document context and previous linking decisions when predicting the current concept.

The target sequence y_t is a textual representation of the gold concept e_t. We construct this target using the same adaptive concept representation strategy as in Chapter 3. Among the unambiguous synonyms of the gold concept, we select the synonym that best matches the mention surface form using TF-IDF similarity. This produces targets that remain close to the observed mentions while still mapping deterministically to unique KB concepts.

2.2.2 Training Objective

The model is trained with teacher forcing to generate y_t from x_t:

$$p_\theta(y_t \mid x_t) = \prod_i p_\theta(y_{t,i} \mid y_{t,<i}, x_t),$$

where $y_{t,i}$ is the i-th token of the target concept string.

2.2.3 Robust Memory Construction with Out-of-Fold Predictions

During inference, the memory $\mathcal{H}_t$ contains the model's own previous predictions, which may be incorrect. If the model is trained only with gold concepts in memory, it may learn to trust this signal too strongly and propagate early errors to later mentions.

To reduce this mismatch between training and inference, we construct the training memory using out-of-fold predictions rather than gold labels. The training set is split into five folds. For each fold, a base model is trained on the other four and used to predict the held-out fold. Each training example therefore receives memory entries generated by a model that was not trained on that example. The final LongBEL model is then trained with these predicted memories. This makes the training memory closer to the memory available at inference time and encourages LongBEL to use previous predictions while remaining robust to their errors. Figure 4.2 illustrates this process.

2.3 Inference

2.3.1 Sequential Prediction with Dynamic Memory

During inference, as illustrated in Figure 4.3, LongBEL resolves mentions sequentially as they appear in the document context c. For the first mention m_1, the historical memory $\mathcal{H}_1$ is initialized empty. Once a prediction e_t is generated for mention m_t, the tuple (m_t, e_t) is appended to update the memory state for the next step, $\mathcal{H}_{t+1}$. This dynamic, autoregressive process explicitly propagates prior linking decisions forward, promoting document-level consistency.

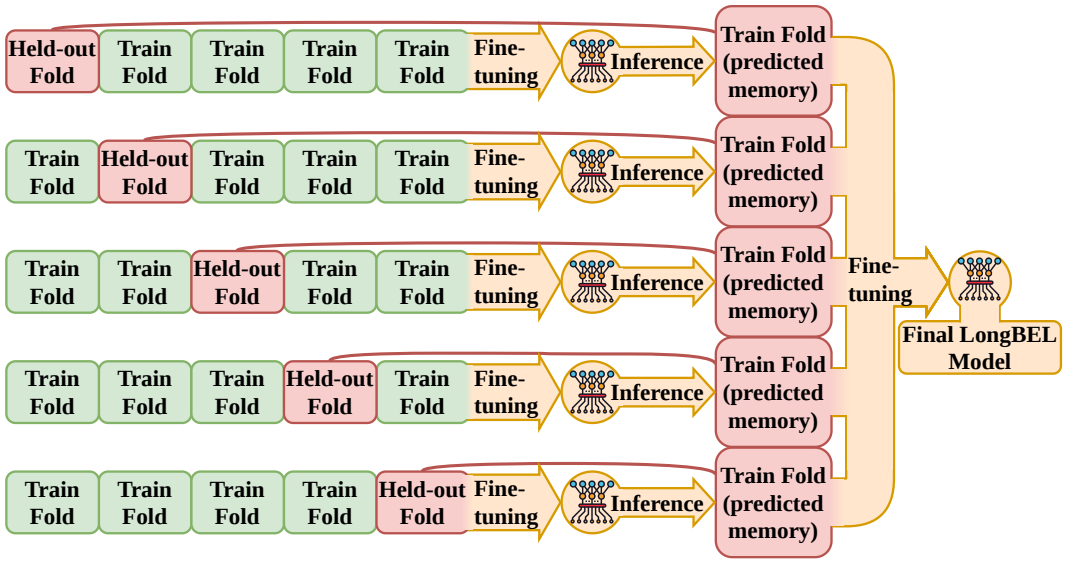


Figure 4.2: Illustration of robust memory construction with out-of-fold predictions. The training set is split into five folds. For each fold, a base model is trained on the remaining folds and predicts the held-out fold. These held-out predictions are then used to create the memory for final LongBEL training, so that the memory used during training better matches inference-time conditions.

2.3.2 Semantic-Guided Constrained Decoding

As in Chapter 3, predictions are constrained to valid KB concepts using semantic-guided constrained decoding. The same inventory of unambiguous synonyms used to construct training targets is used to build the decoding trie. The candidate vocabulary is first restricted to the semantic group of the target mention. Generation is then constrained token by token with this prefix trie, so that each decoded string corresponds to a valid concept from the target KB.

2.3.3 Ensemble via Reciprocal Rank Fusion

In addition to the individual LongBEL configurations, we evaluate an ensemble combining local-context, global-context, memory-only, and global-context-with-memory variants to test whether they provide complementary signals. Each variant produces a ranked list of candidate concepts using constrained beam search. These lists are combined with Reciprocal Rank Fusion (RRF) (Cormack et al., 2009). For each candidate concept e, the RRF score is:

$$\mathrm{RRF}(e) = \sum_{a \in \mathcal{A}} \frac{1}{r_a(e)}$$

where $\mathcal{A}$ is the set of variants and $r_a(e)$ is the rank of e in variant a. Ranks start at 1, and absent candidates receive a score of zero. The candidate with the highest RRF score is selected as the final prediction.

Because the ensemble generates several predictions per mention, it has a higher inference cost than individual configurations.

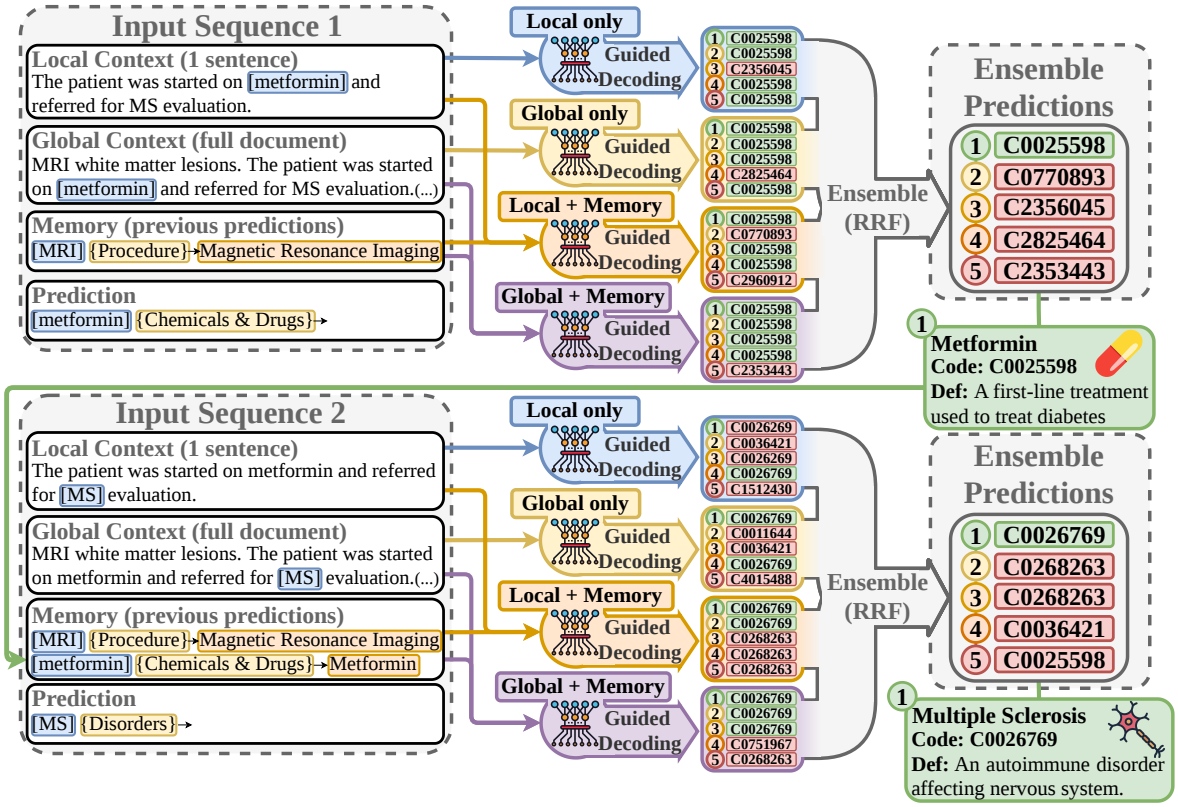


Figure 4.3: Overview of LongBEL inference. For each mention, LongBEL combines local context, full-document context, and memory of previous predictions to generate multiple ranked candidate lists using semantic-guided constrained decoding. These lists are merged with Reciprocal Rank Fusion (RRF) to produce the final prediction. The predicted concept is then added to memory and used for subsequent mentions.

3 Experiments

3.1 Datasets

We reuse the multilingual BEL benchmarks introduced in Chapter 3: **MedMentions-ST21pv (MM-ST21pv)** (Mohan et al., 2019), **QUAERO** (Névéol et al., 2014), and the **Spanish Clinical Case Corpus (SPACCC)** (Miranda-Escalada et al., 2022; Lima-López et al., 2023a; Lima-López et al., 2023b). However, the objective of this chapter is different: we study document-level BEL. We therefore keep only datasets that contain sufficiently long documents. In particular, we exclude QUAERO-MEDLINE, because it consists of short MEDLINE titles and does not provide a meaningful setting for evaluating long-context or memory-based BEL.

Table 4.1 summarizes the datasets used in this chapter. The main addition compared with Chapter 3 is the analysis of concept recurrence within documents. We distinguish *first* mentions, whose concept has not appeared earlier in the document, from *recurring* mentions, whose concept has already been mentioned. MM-ST21pv and QUAERO-EMEA contain many recurring concepts, whereas recurrence is lower in SPACCC.

	MM-ST21pv	QUAERO-EMEA	SPACCC
# Documents	4,392	26	1,000
Average length	1,581	5,062	2,337
Language	English	French	Spanish
Clinical cases	✗	✗	✓
# Mentions	203,282	7,127	27,103
# Concepts	25,419	1,226	9,159
# Sem. Groups	14	10	3
Coverage	1.07%	0.04%	3.34%
% First	48.5%	41.2%	86.1%
% Recurring	51.5%	58.8%	13.9%

Table 4.1: Summary statistics of the evaluation datasets used in this chapter. Average length is reported in characters. Coverage is the proportion of concepts from the target terminology represented in the dataset. % First is the percentage of mentions whose concept appears for the first time in a document. % Recurring is the percentage of mentions whose concept has already appeared at least once earlier in the same document. MM-ST21pv: MedMentions-ST21pv; SPACCC: Spanish Clinical Case Corpus.

3.2 Implementation Details

LongBEL was implemented in Python using PyTorch and the Hugging Face Transformers library. All experiments were run on a single NVIDIA H100 GPU with 80 GB of memory. We reused the same KBs and concept inventories as in Chapter 3, summarized in Table 3.2: Unified Medical Language System (UMLS) 2017AA for MM-ST21pv, UMLS 2014AA for QUAERO-EMEA, and Spanish Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) gazetteers for SPACCC.

The training and inference setup follows Chapter 3, with the same main hyperparameters summarized in Tables 3.5 and 3.6. The only important difference is input length: LongBEL examples can include the full document and the memory of previous predictions, and are therefore longer than the sentence-level examples used in SynCABEL. We therefore adjusted the batch size according to the maximum input length of each dataset.

3.3 Results

3.3.1 Benchmark Performance

Table 4.2 reports Recall@1 on MM-ST21pv, QUAERO-EMEA, and SPACCC. We compare context-free methods, the collective BEL baseline ArboEL, and sentence-level generative models, including SynCABEL. The new methods introduced in this chapter are LongBEL and LongBEL Ensemble. LongBEL uses the full document and previous predictions, while the LongBEL Ensemble combines four separately trained models using local or full-document context, with or without memory.

The most direct comparison is with the corresponding Llama sentence-level baselines, since they use the same backbone. On MM-ST21pv, LongBEL-1B improves from 73.1 to 77.6 and LongBEL-8B from 75.0 to 79.3. The ensemble further improves these

Model	MM-ST21PV (english)	QUAERO-EMEA (french)	SPACCC (spanish)	Average
CONTEXT-FREE BEL				
SciSpacy (Neumann et al., 2019)	53.8±1.0	37.1±4.3	13.2±0.9	34.7
SapBERT (Liu et al., 2021a)	65.6±1.0	59.7±3.8	33.9±1.4	53.1
CODER-all (Yuan et al., 2022c)	62.9±1.1	66.9±4.0	43.7±1.4	57.8
SapBERT-all (Liu et al., 2021b)	64.6±1.1	67.9±3.9	47.9±1.5	60.1
BERGAMOT (Sakhovskiy et al., 2024)	60.9±1.1	63.8±4.9	46.0±1.5	56.9
COLLECTIVE BEL				
ArboEL (Agarwal et al., 2022)	76.9±0.9	63.0±3.9	49.0±1.7	63.0
LOCAL-CONTEXT GENERATIVE BEL				
mBART-large (Tang et al., 2020)	69.6±1.0	69.3±5.4	61.8±2.7	66.9
+ SynCABEL	71.5±1.0	75.3±4.3	64.0±2.3	70.3
Llama-1B (Grattafiori et al., 2024)	73.1±1.0	75.1±3.6	63.8±1.6	70.7
+ SynCABEL	74.2±0.9	75.8±3.0	65.0±1.6	72.4
Llama-8B (Grattafiori et al., 2024)	75.0±0.9	73.8±3.9	65.1±1.5	71.3
+ SynCABEL	75.9±0.8	**79.0**±3.9	<u>67.0</u>±1.3	<u>74.0</u>
GENERATIVE COLLECTIVE BEL				
LongBEL-1B	77.6±0.9	74.5±3.7	63.1±1.6	71.7
LongBEL-1B Ensemble	79.1±0.8	77.2±3.0	65.9±1.5	73.7
LongBEL-8B	<u>79.3</u>±0.8	75.4±3.4	65.3±1.5	73.3
LongBEL-8B Ensemble	**80.6**±0.8	<u>77.6</u>±3.0	**67.2**±1.5	**74.9**

Table 4.2: Entity linking performance (Recall@1) on biomedical benchmarks. The highlighted rows correspond to the LongBEL models introduced in this chapter. The best results are shown in **bold**, the second-best results are <u>underlined</u>, and the "Average" column reports the mean score across the three benchmarks. 95% confidence intervals are estimated using empirical bootstrap resampling at the document level.

results, reaching 79.1 with the 1B models and 80.6 with the 8B models.

On QUAERO-EMEA and SPACCC, the gains are more moderate and depend more on the model variant. The LongBEL-8B Ensemble reaches 77.6 on QUAERO-EMEA and 67.2 on SPACCC, giving the best average score across datasets. SynCABEL remains very competitive, and obtains the best score on QUAERO-EMEA with the 8B backbone. These results suggest that document-level information is useful, but that it complements rather than replaces local context and synthetic augmentation.

Compared with SynCABEL, LongBEL shows a more dataset-dependent behavior. SynCABEL remains best on QUAERO-EMEA, whereas the LongBEL-8B Ensemble obtains the best scores on MM-ST21pv and SPACCC, as well as the best average score. Detailed semantic-group comparisons are reported in Appendix C, Tables S12, S13, S14, and S15.

To ensure a fair comparison, all neural models are evaluated using the same KBs,

Ctx	Mem.	MM-ST21PV	QUAERO-EMEA	SPACCC
		Llama-1B		
✗	✗	73.1±1.0	75.1±3.6	<u>63.8</u>±1.6
✓	✗	76.5±0.9	<u>76.0</u>±2.8	63.7±1.5
✗	✓	77.4±0.8	74.0±3.2	62.8±1.5
✓	✓	<u>77.6</u>±0.9	74.5±3.7	63.1±1.6
Ensemble		**79.1**±0.8	**77.2**±3.0	**65.9**±1.5
		Llama-8B		
✗	✗	75.0±0.9	73.8±4.0	65.1±1.5
✓	✗	78.6±0.9	73.9±3.1	64.9±1.5
✗	✓	79.0±0.9	75.3±3.0	64.4±1.6
✓	✓	<u>79.3</u>±0.8	<u>75.4</u>±3.4	<u>65.3</u>±1.5
Ensemble		**80.6**±0.8	**77.6**±3.0	**67.2**±1.5

Table 4.3: Ablation study (recall) evaluating LongBEL-1B and LongBEL-8B. 'Ctx' denotes the inclusion of global context, and 'Mem.' indicates whether memory is enabled (✓) or disabled (✗). Best results are in **bold**, second-best are <u>underlined</u>. 95% confidence intervals are estimated using empirical bootstrap resampling at the document level. The ensemble rows fuse all components via RRF.

with candidate entities restricted to the semantic group of each mention. The only exception is SciSpacy, whose implementation does not support this filtering.

3.3.2 Ablation Study

To better understand the role of global context (Ctx) and robust memory (Mem), we conduct an ablation study with both LongBEL-1B and LongBEL-8B, as reported in Table 4.3. On MM-ST21pv, both global context and memory help improve performance over the baseline. The best single-model results are obtained when the two components are used together, suggesting that both sources of document-level information are useful on this dataset. Global context allows the model to use evidence from the whole document, while memory gives it access to previous linking decisions that may help maintain consistency across mentions.

On the remaining datasets, improvements are more moderate and less consistent, with differences often within confidence intervals. This was already observed in the main benchmark results and suggests that the usefulness of document-level information varies across benchmarks. We also observe that larger models (8B) integrate global context and memory more reliably, whereas smaller models (1B) exhibit less stable behavior when combining these components.

Finally, in the main Recall@1 comparison, the ensemble obtains the best result on each dataset. This suggests that local context, global context, and memory capture partly different useful signals, and that combining these signals through ensembling leads to the most robust overall performance.

	Llama-1B		Llama-8B	
Subset	Baseline	LongBEL	Baseline	LongBEL
MEDMENTIONS-ST21PV				
Overall	73.1±1.0	77.6±0.9	75.0±0.9	79.3±0.8
First	79.1±0.7	79.1±0.7	80.2±0.7	80.3±0.7
Recurring	67.5±1.4	76.1±1.4	70.1±1.5	78.4±1.3
QUAERO-EMEA				
Overall	75.1±3.6	74.5±3.7	73.8±4.0	75.4±3.4
First	75.0±3.4	72.6±3.6	73.6±3.2	74.2±3.5
Recurring	75.2±5.4	75.8±5.2	74.0±5.6	76.3±4.2
SPACCC				
Overall	63.8±1.6	63.1±1.6	65.1±1.5	65.3±1.5
First	62.2±1.6	61.4±1.6	63.7±1.6	63.4±1.5
Recurring	73.4±3.1	73.3±3.0	74.0±3.0	76.8±2.8

Table 4.4: Recall comparison across datasets, broken down by concept frequency within documents for both the 1B and 8B models. "Baseline" corresponds to the underlying Llama models applied at the sentence level without cross-mention memory. "LongBEL" extends this model with global document context and robust prediction memory. "First" denotes mentions whose concept appears for the first time in the document, while "Recurring" denotes mentions whose concept has already appeared earlier in the document. 95% confidence intervals were estimated via the empirical bootstrap method at the document level.

4 Discussion

4.1 Impact of Concept Redundancy

The improvements brought by LongBEL mainly come from its ability to enforce consistency across concepts that recur within a document. In Table 4.4, we use *recurring concepts* to refer to concepts that have already appeared earlier in the same document. Gains are mainly concentrated on these recurring concepts, while performance on first occurrences stays similar or slightly decreases. For example, on MM-ST21pv, performance on recurring concepts increases from 70.1 to 78.4 for the 8B model, while performance on first occurrences remains nearly unchanged. This suggests that LongBEL does not mainly improve local disambiguation, which is already well handled by sentence-level models. Instead, it helps by sharing information across mentions of the same concept. This is especially useful for repeated concepts, where earlier mentions can disambiguate later underspecified forms.

This explains the dataset-dependent trends observed in the results. The effectiveness of LongBEL depends on how often concepts recur within a document. Datasets with a high proportion of recurring concepts, such as MM-ST21pv and QUAERO-EMEA (51.5% and 58.8%, respectively), provide more opportunities for LongBEL to exploit consistency signals. In contrast, the SPACCC clinical datasets contain far fewer recurring concepts (around 10–16%), which limits the model's impact. Nevertheless, QUAERO-

EMEA shows only moderate gains despite its high redundancy rate, suggesting that recurrence alone is not sufficient.

The semantic group analysis in Appendix C.2, Figures S3a and S3b, helps explain these differences. On MM-ST21pv, recurring concepts are numerous and LongBEL produces stable gains, with an overall improvement of $+8.6$ Recall@1 points for the 1B model and $+8.2$ points for the 8B model. QUAERO-EMEA is different. Although many mentions correspond to recurring concepts, there are only 1,257 recurring mentions in total, and some semantic groups contain very few examples. For instance, *Phenomena* has only 9 recurring mentions and shows an unstable effect, with about -11 points for LongBEL-1B and approximately no gain for LongBEL-8B. Similarly, *Objects* has only 22 recurring mentions, with approximately no gain for LongBEL-1B and a negative delta of about -14 points for LongBEL-8B. These small supports make group-level estimates less stable, even when recurrence exists. This suggests that LongBEL does not only require concepts to recur within documents. It also needs enough recurring examples within each semantic group to exploit consistency signals reliably.

Overall, these results show that LongBEL acts as a document-level consistency mechanism. Its effectiveness depends on whether concepts recur within a document and is influenced by both concept redundancy and sample size, rather than by improvements on isolated mentions.

4.2 Detailed Subset Analysis

Tables 4.5–4.7 extend the previous analysis by reporting performance according to concept familiarity, lexical match, mention length, and document recurrence. The support column gives the percentage and number of examples in each subset. The first two rows report overall Recall@1 and Recall@5, while the remaining rows report Recall@1.

The clearest trends are observed on MM-ST21pv. LongBEL-8B improves substantially on unseen concepts, from 47.1 to 59.2, and on non-identical mentions, from 73.8 to 78.3. However, its largest gains remain concentrated on recurring concepts: performance increases from 70.1 to 78.4, while remaining almost unchanged on first occurrences. The gains also increase with recurrence rank, reaching 14.6 points for mentions occurring at least five times.

The trends are less stable on QUAERO-EMEA and SPACCC because several subsets contain fewer examples and have wider confidence intervals. Nevertheless, LongBEL-8B again improves more on recurring than on first occurrences. Other properties, such as lexical match and mention length, do not show a consistent effect across datasets. These results confirm that the main contribution of LongBEL comes from exploiting repeated concepts within documents rather than uniformly improving all types of mentions.

4.3 Qualitative Examples

Tables 4.8 and 4.9 provide representative comparisons between LongBEL-8B and the sentence-level Llama-8B baseline. The first table illustrates cases where document-level information recovers the correct concept. The second illustrates error propagation, a case where document-level evidence remains insufficient, and a simple case where both models succeed.

Subset	Support	1B			8B		
		Base	**LongBEL**	**Ens.**	**Base**	**LongBEL**	**Ens.**
Overall performance							
Recall@1 overall	100% (40,143)	$73.1_{\pm 1.0}$	$77.6_{\pm 0.9}$	$79.1_{\pm 0.9}$	$75.0_{\pm 0.9}$	$79.3_{\pm 0.8}$	$80.6_{\pm 0.8}$
Recall@5 overall	100% (40,143)	$81.2_{\pm 0.8}$	$85.8_{\pm 0.7}$	$85.1_{\pm 0.7}$	$83.7_{\pm 0.9}$	$86.5_{\pm 0.8}$	$86.3_{\pm 0.7}$
Concept familiarity							
Seen concepts	79.5% (31,914)	$80.5_{\pm 0.9}$	$82.8_{\pm 0.8}$	$84.5_{\pm 0.8}$	$82.2_{\pm 0.8}$	$84.4_{\pm 0.8}$	$85.8_{\pm 0.7}$
Unseen concepts	20.5% (8,229)	$44.6_{\pm 2.1}$	$57.5_{\pm 2.4}$	$58.3_{\pm 2.5}$	$47.1_{\pm 2.2}$	$59.2_{\pm 2.5}$	$60.2_{\pm 2.5}$
Lexical match							
Identical	5.6% (2,265)	$93.5_{\pm 1.9}$	$95.5_{\pm 1.3}$	$95.9_{\pm 1.4}$	$94.6_{\pm 1.4}$	$95.2_{\pm 1.4}$	$95.6_{\pm 1.4}$
Not identical	94.4% (37,878)	$71.9_{\pm 0.9}$	$76.5_{\pm 0.9}$	$78.1_{\pm 0.9}$	$73.8_{\pm 0.9}$	$78.3_{\pm 0.8}$	$79.7_{\pm 0.9}$
Mention length							
Single word	66.8% (26,833)	$74.9_{\pm 1.1}$	$80.4_{\pm 1.0}$	$81.8_{\pm 1.0}$	$76.6_{\pm 1.1}$	$82.3_{\pm 1.0}$	$83.2_{\pm 0.9}$
Multi-word	33.2% (13,310)	$69.5_{\pm 1.2}$	$72.0_{\pm 1.2}$	$73.7_{\pm 1.1}$	$71.7_{\pm 1.1}$	$73.2_{\pm 1.1}$	$75.2_{\pm 1.2}$
Document recurrence							
First occurrence	48.5% (19,487)	$79.1_{\pm 0.7}$	$79.1_{\pm 0.7}$	$80.8_{\pm 0.7}$	$80.2_{\pm 0.6}$	$80.3_{\pm 0.7}$	$82.0_{\pm 0.6}$
Recurring occurrence	51.5% (20,656)	$67.5_{\pm 1.4}$	$76.1_{\pm 1.4}$	$77.5_{\pm 1.3}$	$70.1_{\pm 1.5}$	$78.4_{\pm 1.3}$	$79.3_{\pm 1.2}$
2nd occurrence	18.5% (7,436)	$75.8_{\pm 1.1}$	$78.6_{\pm 1.0}$	$80.1_{\pm 0.9}$	$77.1_{\pm 1.0}$	$80.2_{\pm 0.9}$	$81.4_{\pm 0.9}$
3rd occurrence	10.4% (4,167)	$69.6_{\pm 1.6}$	$77.1_{\pm 1.4}$	$78.7_{\pm 1.3}$	$72.2_{\pm 1.6}$	$79.0_{\pm 1.4}$	$80.0_{\pm 1.2}$
4th occurrence	6.7% (2,704)	$65.0_{\pm 1.9}$	$75.8_{\pm 1.7}$	$77.2_{\pm 1.6}$	$68.2_{\pm 1.8}$	$77.5_{\pm 1.5}$	$78.6_{\pm 1.5}$
5th+ occurrence	15.8% (6,349)	$57.3_{\pm 2.8}$	$72.7_{\pm 2.5}$	$73.9_{\pm 2.5}$	$61.4_{\pm 2.6}$	$76.0_{\pm 2.4}$	$76.5_{\pm 2.3}$

Table 4.5: Detailed subset results on MM-ST21pv. The first two rows report overall Recall@1 and Recall@5; all other rows report Recall@1 by subset. Support gives the percentage and number of examples. Scores are reported with half of the 95% confidence interval. Base denotes the baseline model and Ens. the ensemble.

Subset	Support	1B			8B		
		Base	**LongBEL**	**Ens.**	**Base**	**LongBEL**	**Ens.**
Overall performance							
Recall@1 overall	100% (2,196)	$75.1_{\pm 3.6}$	$74.5_{\pm 3.7}$	$77.2_{\pm 3.0}$	$73.8_{\pm 4.0}$	$75.4_{\pm 3.4}$	$77.6_{\pm 3.0}$
Recall@5 overall	100% (2,196)	$82.0_{\pm 3.0}$	$82.4_{\pm 2.8}$	$82.8_{\pm 3.2}$	$82.2_{\pm 2.6}$	$82.0_{\pm 2.5}$	$82.8_{\pm 2.7}$
Concept familiarity							
Seen concepts	56.0% (1,229)	$88.7_{\pm 2.7}$	$89.3_{\pm 2.1}$	$91.1_{\pm 1.8}$	$92.1_{\pm 2.1}$	$90.8_{\pm 2.8}$	$92.8_{\pm 1.2}$
Unseen concepts	44.0% (967)	$57.5_{\pm 5.8}$	$55.5_{\pm 5.5}$	$59.2_{\pm 5.6}$	$50.2_{\pm 5.1}$	$55.8_{\pm 5.4}$	$58.0_{\pm 5.6}$
Lexical match							
Identical	6.5% (143)	$93.7_{\pm 4.4}$	$97.9_{\pm 2.5}$	$100.0_{\pm 0.0}$	$93.0_{\pm 4.6}$	$99.3_{\pm 1.0}$	$98.6_{\pm 1.8}$
Not identical	93.5% (2,053)	$73.8_{\pm 3.4}$	$72.8_{\pm 3.9}$	$75.6_{\pm 3.1}$	$72.5_{\pm 4.0}$	$73.7_{\pm 3.6}$	$76.2_{\pm 2.9}$
Mention length							
Single word	77.6% (1,704)	$79.9_{\pm 4.0}$	$80.8_{\pm 4.0}$	$82.1_{\pm 3.5}$	$78.5_{\pm 4.8}$	$80.3_{\pm 4.9}$	$82.0_{\pm 3.7}$
Multi-word	22.4% (492)	$58.5_{\pm 6.4}$	$52.6_{\pm 9.1}$	$60.2_{\pm 8.2}$	$57.7_{\pm 5.7}$	$58.3_{\pm 6.0}$	$62.4_{\pm 5.6}$
Document recurrence							
First occurrence	42.8% (939)	$75.0_{\pm 3.4}$	$72.6_{\pm 3.6}$	$75.6_{\pm 3.5}$	$73.6_{\pm 3.1}$	$74.2_{\pm 3.5}$	$76.3_{\pm 2.8}$
Recurring occurrence	57.2% (1,257)	$75.2_{\pm 5.4}$	$75.8_{\pm 5.1}$	$78.4_{\pm 4.6}$	$74.0_{\pm 5.5}$	$76.3_{\pm 4.9}$	$78.7_{\pm 4.3}$
2nd occurrence	18.9% (416)	$77.4_{\pm 2.7}$	$76.0_{\pm 4.6}$	$79.1_{\pm 3.3}$	$76.2_{\pm 4.4}$	$77.6_{\pm 3.2}$	$79.3_{\pm 3.2}$
3rd occurrence	9.6% (210)	$72.4_{\pm 5.7}$	$71.4_{\pm 5.4}$	$75.7_{\pm 4.1}$	$72.9_{\pm 4.4}$	$74.8_{\pm 3.5}$	$75.2_{\pm 4.2}$
4th occurrence	5.5% (120)	$77.5_{\pm 5.3}$	$76.7_{\pm 5.0}$	$78.3_{\pm 4.8}$	$75.8_{\pm 6.0}$	$76.7_{\pm 5.5}$	$79.2_{\pm 4.4}$
5th+ occurrence	23.3% (511)	$74.0_{\pm 9.6}$	$77.3_{\pm 9.2}$	$78.9_{\pm 8.6}$	$72.2_{\pm 10.1}$	$75.7_{\pm 9.7}$	$79.5_{\pm 8.2}$

Table 4.6: Detailed subset results on QUAERO-EMEA. The first two rows report overall Recall@1 and Recall@5; all other rows report Recall@1 by subset. Support gives the percentage and number of examples. Scores are reported with half of the 95% confidence interval. Base denotes the baseline model and Ens. the ensemble.

Subset	Support	1B			8B		
		Base	LongBEL	Ens.	Base	LongBEL	Ens.
Overall performance							
Recall@1 overall	100% (8,821)	$63.8_{\pm 1.6}$	$63.1_{\pm 1.6}$	$65.9_{\pm 1.5}$	$65.1_{\pm 1.5}$	$65.3_{\pm 1.5}$	$67.2_{\pm 1.5}$
Recall@5 overall	100% (8,821)	$70.4_{\pm 1.4}$	$69.1_{\pm 1.5}$	$74.2_{\pm 1.3}$	$72.1_{\pm 1.4}$	$71.4_{\pm 1.4}$	$73.9_{\pm 1.5}$
Concept familiarity							
Seen concepts	69.7% (6,145)	$82.0_{\pm 1.3}$	$80.8_{\pm 1.2}$	$83.7_{\pm 1.2}$	$83.4_{\pm 1.2}$	$83.6_{\pm 1.2}$	$85.3_{\pm 1.1}$
Unseen concepts	30.3% (2,676)	$22.0_{\pm 1.8}$	$22.5_{\pm 1.7}$	$25.0_{\pm 1.8}$	$23.2_{\pm 1.8}$	$23.3_{\pm 1.8}$	$25.6_{\pm 2.0}$
Lexical match							
Identical	21.2% (1,868)	$96.0_{\pm 0.9}$	$94.8_{\pm 1.2}$	$97.2_{\pm 0.9}$	$95.8_{\pm 1.0}$	$96.1_{\pm 1.0}$	$96.9_{\pm 0.8}$
Not identical	78.8% (6,953)	$55.1_{\pm 1.7}$	$54.6_{\pm 1.8}$	$57.5_{\pm 1.7}$	$56.9_{\pm 1.7}$	$57.0_{\pm 1.7}$	$59.2_{\pm 1.6}$
Mention length							
Single word	30.2% (2,665)	$87.8_{\pm 1.8}$	$86.8_{\pm 1.8}$	$89.0_{\pm 1.5}$	$88.7_{\pm 1.6}$	$89.1_{\pm 1.6}$	$89.7_{\pm 1.7}$
Multi-word	69.8% (6,156)	$53.4_{\pm 1.8}$	$52.8_{\pm 1.9}$	$55.9_{\pm 1.8}$	$55.0_{\pm 1.8}$	$55.0_{\pm 1.8}$	$57.4_{\pm 1.8}$
Document recurrence							
First occurrence	85.7% (7,559)	$62.2_{\pm 1.6}$	$61.4_{\pm 1.6}$	$64.1_{\pm 1.6}$	$63.7_{\pm 1.6}$	$63.4_{\pm 1.6}$	$65.3_{\pm 1.6}$
Recurring occurrence	14.3% (1,262)	$73.4_{\pm 3.1}$	$73.3_{\pm 3.0}$	$76.5_{\pm 2.8}$	$74.0_{\pm 3.0}$	$76.8_{\pm 2.8}$	$78.1_{\pm 3.0}$
2nd occurrence	10.5% (930)	$71.6_{\pm 3.0}$	$70.9_{\pm 3.1}$	$74.5_{\pm 3.0}$	$72.2_{\pm 3.0}$	$73.3_{\pm 2.8}$	$75.5_{\pm 2.9}$
3rd occurrence	2.4% (213)	$78.4_{\pm 5.2}$	$78.4_{\pm 5.4}$	$80.8_{\pm 4.8}$	$76.5_{\pm 5.4}$	$84.5_{\pm 4.6}$	$83.6_{\pm 5.0}$
4th occurrence	0.8% (68)	$75.0_{\pm 10.4}$	$79.4_{\pm 10.2}$	$80.9_{\pm 9.6}$	$82.4_{\pm 10.0}$	$85.3_{\pm 8.6}$	$86.8_{\pm 9.3}$
5th+ occurrence	0.6% (51)	$82.4_{\pm 10.4}$	$88.2_{\pm 10.4}$	$88.2_{\pm 10.0}$	$86.3_{\pm 12.4}$	$96.1_{\pm 6.6}$	$92.2_{\pm 9.0}$
Concept structure							
Simple concepts	96.3% (8,495)	$65.4_{\pm 1.6}$	$64.6_{\pm 1.6}$	$67.6_{\pm 1.5}$	$66.8_{\pm 1.6}$	$67.1_{\pm 1.5}$	$68.7_{\pm 1.5}$
Composite concepts	3.7% (326)	$22.1_{\pm 4.8}$	$23.0_{\pm 4.8}$	$21.2_{\pm 4.6}$	$22.4_{\pm 4.4}$	$19.0_{\pm 3.8}$	$26.1_{\pm 4.8}$

Table 4.7: Detailed subset results on SPACCC. The first two rows report overall Recall@1 and Recall@5; all other rows report Recall@1 by subset. Support gives the percentage and number of examples. Scores are reported with half of the 95% confidence interval. Base denotes the baseline model and Ens. the ensemble. Simple and composite concept rows are reported only for SPACCC.

Outcome	Example
LongBEL ✓ **Baseline** ✗	**Mention:** [ND] **Local context:** Therefore, we investigated factors associated with [ND] in patients with LAA. **Relevant global context:** In some patients with acute ischemic stroke, neurological deterioration (ND) is observed, and it is difficult to predict at the time of admission. **Previous normalized mentions:** neurological deterioration → Neurological status deteriorated ND → Neurological status deteriorated large-artery atherosclerosis → Large artery atherosclerosis **Gold:** Neurological status deteriorated (C1536136) **Baseline:** Nicotine dependence (disorder) (C0028043) ✗ **LongBEL-8B:** Neurological status deteriorated (C1536136) ✓ *Takeaway:* The abbreviation "ND" is explicitly defined earlier in the document as neurological deterioration. The sentence-level baseline only sees the ambiguous abbreviation and links it to nicotine dependence, whereas LongBEL uses the previous abbreviation definition to recover the correct concept.
LongBEL ✓ **Baseline** ✗	**Mention:** [PET] **Local context:** She was admitted overnight with suspected [PET] and was started on urgent treatment. **Relevant global context:** A 29-year-old pregnant woman presented with severe-range hypertension, headache, and epigastric pain. Laboratory testing showed proteinuria and mildly elevated liver enzymes. **Previous normalized mentions:** pregnant woman → Pregnant Woman severe-range hypertension → Hypertensive disease proteinuria → Proteinurias **Gold:** Pre-eclamptic toxemia (C0032914) **Baseline:** Petechial haemorrhage (C0031256) ✗ **LongBEL-8B:** Pre-eclamptic toxemia (C0032914) ✓ *Takeaway:* Unlike an explicit abbreviation-definition case, this example requires combining several document-level clinical cues. The local sentence only contains the ambiguous abbreviation "PET", while LongBEL uses the obstetric context and previous normalized mentions, especially "pregnant Woman", "hypertensive disease", and "proteinurias", to recover the correct concept.

Table 4.8: Qualitative examples where document-level context helps LongBEL-8B. The first example shows an explicit abbreviation definition, while the second shows a case requiring broader clinical context. The ND example is from MedMentions-ST21pv; the PET example is manually constructed.

Outcome	Example
LongBEL ✗	**Mention:** [IFN-γ] **Local context:** The aim of the study is to investigate the association of [IFN-γ] and IL-10 gene single nucleotide polymorphisms.
Baseline ✓	**Global context:** (...) Interferon gamma (IFN-γ) and interleukin 10 (IL-10) polymorphisms in Chinese children (...) **Previous normalized mentions:** Interferon gamma $\rightarrow$ Interleukin 9 Gene IFN-γ $\rightarrow$ Interleukin 9 Gene interleukin 10 $\rightarrow$ Interleukin 10 IL-10 $\rightarrow$ Interleukin 10 **Gold:** IFNG gene (C1334085) **Baseline:** IFNG gene (C1334085) ✓ **LongBEL-8B:** Interleukin 9 Gene (C1334127) ✗ *Takeaway:* Although the document contains an explicit definition, LongBEL mislinks the first occurrence "Interferon gamma (IFN-γ)" and then propagates this incorrect prediction through memory to the later mention.
LongBEL ✗	**Mention:** [MK-8(H2)] **Local context:** The major menaquinone were [MK-8(H2)] (72%) and MK-9(H2) (28%).
Baseline ✗	**Gold:** vitamin MK 8 (C0084980) **Baseline:** MAPKAPK2 protein, human (C1452441) ✗ **LongBEL-8B:** vitamin K2 (C0086605) ✗ *Takeaway:* The mention is a specialized chemical abbreviation with no additional disambiguating evidence elsewhere in the document. LongBEL predicts a related vitamin K concept, but both models fail to recover the exact KB concept.
LongBEL ✓	**Mention:** [DDIs] **Local context:** Elderly patients are at high risk from drug-drug interactions ([DDIs]).
Baseline ✓	**Gold:** Drug Interactions (C0687133) **Baseline:** Drug Interactions (C0687133) ✓ **LongBEL-8B:** Drug Interactions (C0687133) ✓ *Takeaway:* The local sentence explicitly defines the abbreviation, so both the sentence-level baseline and LongBEL resolve the mention correctly.

Table 4.9: Qualitative examples where document-level memory does not improve over the baseline. The examples illustrate error propagation, failure on a specialized chemical abbreviation, and a simple case where both models succeed. All examples are from MedMentions-ST21pv.

4.4 Robust Memory Mitigates Cascading Errors

LongBEL relies on previous predictions, which creates a risk of error propagation. If the model is trained with gold memory, it may learn to trust this signal too strongly: an early incorrect prediction can then be copied to later mentions of the same concept. To reduce this mismatch between training and inference, LongBEL uses robust memory built from out-of-fold model predictions, as described in Section 2.2.3.

We evaluate this effect with the *Copy Wrong Memory Error Rate* (CWME). This metric measures how often the model repeats an earlier wrong prediction for later mentions of the same gold concept. Lower values therefore indicate fewer cascading errors. With robust memory, CWME decreases for both model sizes: by up to 20.7 percentage points on MM-ST21pv, 4.5 points on QUAERO-EMEA, and 10.5 points on SPACCC. Robust memory also improves overall linking performance. For LongBEL-8B, Recall@1 increases from 78.1 to 79.3 on MM-ST21pv, from 75.2 to 75.4 on QUAERO-EMEA, and from 63.9 to 65.3 on SPACCC. These results suggest that exposing the model to predicted memory during training helps it use previous decisions more cautiously, reducing cascading errors without sacrificing overall accuracy. The formal definition of CWME and the detailed results by semantic group are provided in Appendix C, Figure S4.

4.5 Generalization Across Datasets

Table 4.10 evaluates transfer between different training and evaluation corpora. All models use the same Llama-3-8B backbone. Across the six available cross-dataset comparisons, LongBEL consistently outperforms the sentence-level Baseline, with gains ranging from 5.4 to 15.6 Recall@1 points. This shows that document-level context and memory remain useful beyond the corpus used for training. Nevertheless, performance is still lower than in the corresponding within-dataset settings, showing that transfer between corpora remains difficult.

The absolute transfer results also differ across target datasets. On MM-ST21pv, cross-dataset models generally perform in the same range as context-free neural methods. On SPACCC, they remain below most context-free baselines. In contrast, several transfer configurations on QUAERO-EMEA and QUAERO-MEDLINE remain competitive with, or outperform, context-free methods. This suggests that cross-dataset supervision can still provide useful contextual information, although its value depends strongly on the target corpus. This difficulty is especially clear for transfers involving SPACCC. Unlike MM-ST21pv and QUAERO-EMEA, SPACCC uses SNOMED CT rather than UMLS, follows a different annotation scheme, contains generally longer mentions, and consists of clinical case reports. These differences create a larger shift in language, mention boundaries, concept space, and document structure.

SynCABEL shows a less consistent transfer pattern. Its clearest gains occur when it is trained on QUAERO-EMEA and evaluated on MM-ST21pv or QUAERO-MEDLINE, while performance decreases in most other cross-dataset settings. This suggests that synthetic augmentation can improve coverage of the source concept space, but may transfer poorly when the target corpus uses a different KB or annotation scheme. Overall, LongBEL provides a more stable cross-dataset benefit, whereas the effect of SynCABEL depends more strongly on the source and target corpora.

Training dataset	Model	Evaluation dataset			
		MM-ST21pv	QUAERO-EMEA	QUAERO-MEDLINE	SPACCC
MM-ST21pv	Baseline	75.0±0.9	62.3±4.6	**70.3±1.6**	16.2±1.0
	SynCABEL	75.9±0.8	55.6±6.3	67.2±1.8	14.4±1.0
	LongBEL	**79.3±0.8**	**72.4±2.7**	—	**25.6±1.4**
QUAERO-EMEA	Baseline	56.4±1.0	73.8±4.0	65.7±1.8	11.8±0.9
	SynCABEL	**62.7±1.0**	**79.0±3.9**	**74.2±1.6**	5.3±0.5
	LongBEL	61.8±1.1	75.4±3.4	—	**27.4±1.6**
QUAERO-MEDLINE	Baseline	60.7±1.0	**72.6±3.0**	77.6±1.6	**14.2±0.9**
	SynCABEL	**61.0±1.0**	71.8±4.4	**79.7±1.4**	12.8±0.9
	LongBEL	—	—	—	—
SPACCC	Baseline	43.6±1.0	41.4±6.0	**41.4±1.9**	65.1±1.5
	SynCABEL	20.1±0.8	16.4±5.2	35.5±1.8	**67.0±1.3**
	LongBEL	**56.8±1.1**	**56.4±3.3**	—	65.3±1.5

Table 4.10: Cross-dataset biomedical entity linking performance in Recall@1 (%). Each model is trained on the dataset indicated in the rows and evaluated, without further training, on the dataset indicated in the columns. All models use the same Llama-3-8B backbone. Values are reported as the point estimate ± half of the 95% confidence interval. The best score for each training–evaluation dataset pair is shown in bold. Dashes indicate unavailable configurations.

4.6 Saliency Maps Illustrate the Use of Document-Level Evidence

To further examine the successful predictions presented above, we complement the qualitative examples with token-level saliency maps. These maps indicate which input tokens the generated prediction is most sensitive to. They are not intended as a causal explanation of the model, but as a simple way to inspect whether LongBEL uses document context and memory entries when selecting a concept.

We compute saliency with a gradient-based method, following the general idea of saliency maps introduced by Simonyan et al. (2014). More broadly, this analysis relates to work on explanations for natural language processing models (Nguyen, 2024). After decoding, we keep the generated output sequence y and define the score:

$$S(x, y) = \sum_{t \in y} \log p_\theta(y_t \mid x, y_{<t}),$$

where x is the input prompt and y is the generated concept name. We then compute the gradient of this score with respect to each input token embedding. The saliency score of token i is the ℓ_2 norm of this gradient:

$$a_i = \left\| \frac{\partial S(x, y)}{\partial e_i} \right\|_2,$$

where e_i is the embedding of the i-th input token. Scores are normalized by the maximum score within each example. We only display saliency for input prompt tokens.

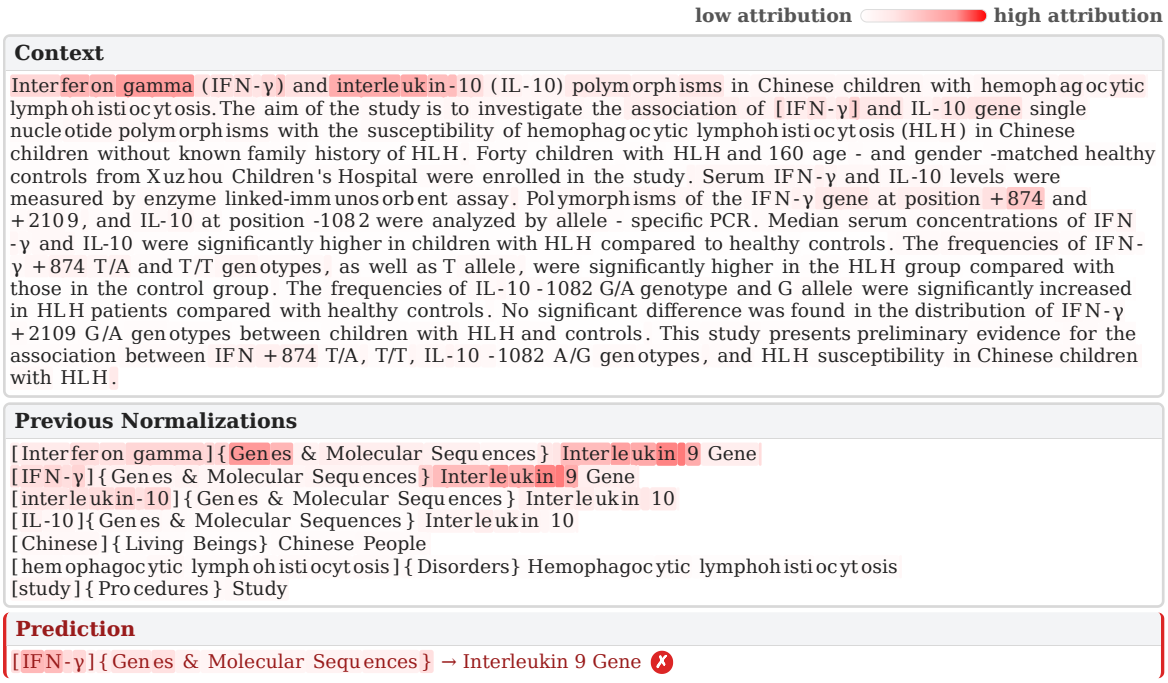


Figure 4.4: Saliency maps for LongBEL-8B predictions. Failure case from MedMentions-ST21pv. In the IFN-γ example, LongBEL-8B gives high saliency to an incorrect previous prediction, showing that an early linking error can influence a later abbreviation prediction through memory.

Figure 4.4 first illustrates a failure case. LongBEL gives high saliency to an incorrect memory entry for *Interferon gamma*, showing how an early prediction error can influence later predictions of the same concept.

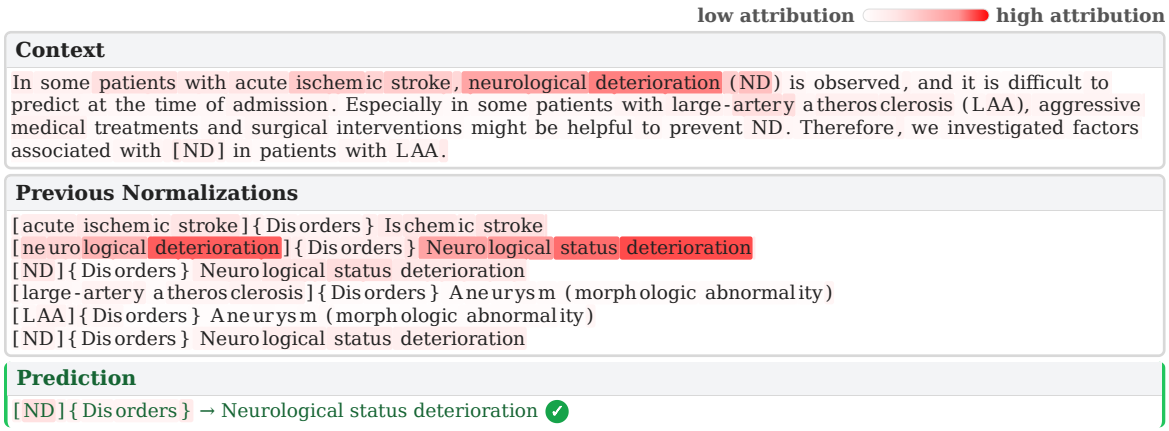
Figures 4.5a and 4.5b show the opposite behavior, where document-level evidence helps LongBEL recover the correct concept. In the ND example, the model assigns high saliency to the earlier definition “neurological deterioration (ND)” and to the corresponding memory entries. In the PET example, LongBEL relies on broader clinical context and previous linked mentions, such as pregnancy, hypertension, and proteinuria, to interpret *PET* as pre-eclamptic toxemia.

Together, these examples support the interpretation of LongBEL as a document-level consistency mechanism. Memory and global context can help resolve ambiguous mentions, but they can also transmit earlier mistakes when the memory is incorrect.

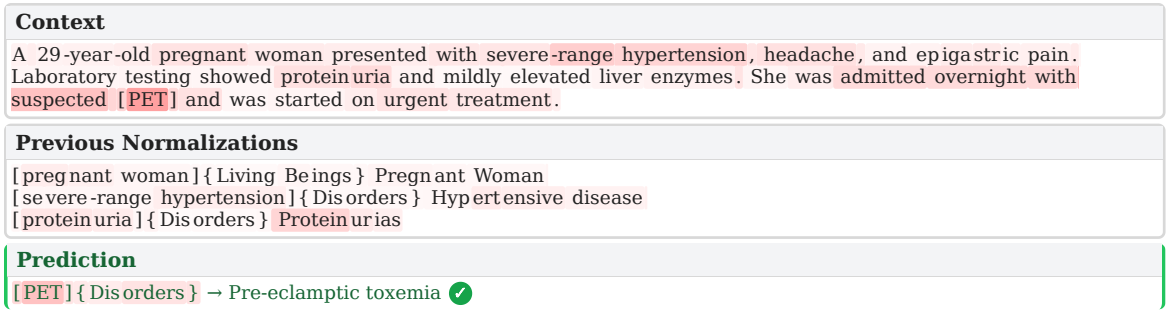
4.7 Efficiency Trade-offs

The inference footprint of the main BEL model families was already discussed in Chapter 3. Here, we extend this analysis to LongBEL to measure the cost of document-level modeling. Table 4.11 reports the same indicators as before, while adding the LongBEL variants.

LongBEL does not increase memory requirements. For a given backbone, it uses the same model parameters and the same candidate trie as the corresponding sentence-level SynCABEL model. The additional cost is inference time. With the 1B backbone, throughput decreases from 69.6 to 48.5 mentions/s. With the 8B backbone, it decreases



(a) Abbreviation-resolution example from MedMentions-ST21pv.



(b) Manually constructed example involving clinical context and memory.

Figure 4.5: Saliency maps for LongBEL-8B predictions. These two examples illustrate cases where document-level evidence helps. In the ND example, LongBEL gives high saliency to the earlier definition “neurological deterioration (ND)” and to the corresponding memory entries. In the PET example, it uses obstetric context and previous linked mentions, especially hypertension and proteinuria, to link PET to pre-eclamptic toxemia.

Model	Model size (GB)	Candidate size (GB)	Speed (mentions/s)
SapBERT	2.1	20.1	**575.5**
ArboEL	**1.2**	7.1	38.9
mBART	2.3	**5.4**	51.0
SynCABEL-1B	2.4	**5.4**	69.6
LongBEL-1B	2.4	**5.4**	48.5
LongBEL-1B-Ensemble	2.4	**5.4**	14.8
SynCABEL-8B	28.6	**5.4**	38.2
LongBEL-8B	28.6	**5.4**	15.2
LongBEL-8B-Ensemble	28.6	**5.4**	6.7

Table 4.11: Inference speed and memory footprint. Model size refers to the memory required by model parameters, candidate size refers to the memory required by the candidate representation, and speed is reported in mentions per second.

from 38.2 to 15.2 mentions/s. This slowdown comes from the longer inputs and the sequential update of the prediction memory.

These results highlight LongBEL's main trade-off. It improves document-level consistency without adding parameters or candidate storage, but slows inference. SynCABEL remain preferable when throughput is the main constraint, whereas LongBEL is more relevant when document-level coherence matters.

4.8 End-to-End Impact of Document-Level Normalization

We evaluate LongBEL in the same end-to-end setting as in Chapter 3. The named entity recognition (NER) predictions are fixed for all models, so differences between CODER-all, SapBERT-all, SynCABEL, and LongBEL reflect only the normalization step. A prediction is correct only when both the detected span and the normalized concept are correct. For threshold selection, we use the same validation procedure as in Chapter 3. Each normalization model uses a confidence threshold to accept or reject predictions. The threshold is selected by maximizing end-to-end F1-score on the validation set over 1,000 candidate thresholds sampled from the empirical score distribution, then applied unchanged to the test set. The corresponding curves are reported in Figure S5 in Appendix C.

Table 4.12 shows that LongBEL improves the end-to-end pipeline on MM-ST21pv, where recurring concepts are frequent. It increases F1-score from 55.4 with SynCABEL to 58.2, mainly through higher recall. On QUAERO-EMEA, SynCABEL remains better than LongBEL, with an F1-score of 64.2 compared with 61.5. This is consistent with the previous results, where LongBEL was less stable on this dataset. On SPACCC, both methods are very close: LongBEL slightly improves precision, while SynCABEL keeps a small advantage in recall and F1-score.

Overall, LongBEL is useful when document-level consistency signals are strong, as in MM-ST21pv, but it does not systematically replace SynCABEL. The two approaches remain complementary: SynCABEL improves concept coverage through synthetic examples, whereas LongBEL exploits repeated evidence within a document.

Dataset	Pipeline setting	Precision	Recall	F1-score
MM-ST21pv	*NER only* (He et al., 2022)	69.0 ± 0.9	67.9 ± 0.9	68.5 ± 0.8
	NER + CODER-all	60.5 ± 1.0	43.7 ± 0.9	50.7 ± 0.9
	NER + SapBERT-all	58.9 ± 1.1	42.0 ± 1.0	49.0 ± 1.0
	NER + SynCABEL	61.0 ± 1.0	50.7 ± 1.0	55.4 ± 0.9
	NER + LongBEL	**62.2** ± 1.0	**54.7** ± 0.9	**58.2** ± 0.9
QUAERO-EMEA	*NER only* (Touchent et al., 2024)	79.9 ± 2.9	65.3 ± 5.5	71.9 ± 3.9
	NER + CODER-all	72.2 ± 4.6	44.9 ± 5.6	55.4 ± 5.2
	NER + SapBERT-all	60.6 ± 3.8	44.2 ± 5.0	51.1 ± 4.1
	NER + SynCABEL	**74.5** ± 2.7	**56.4** ± 5.9	**64.2** ± 4.3
	NER + LongBEL	73.7 ± 3.9	52.7 ± 5.8	61.5 ± 4.6
SPACCC	*NER only* (Carrino et al., 2022)	76.3 ± 1.4	57.6 ± 2.0	65.7 ± 1.6
	NER + CODER-all	59.2 ± 1.7	26.7 ± 1.4	36.8 ± 1.6
	NER + SapBERT-all	55.3 ± 1.8	28.3 ± 1.5	37.4 ± 1.6
	NER + SynCABEL	70.5 ± 1.9	**38.3** ± 1.9	**49.7** ± 1.9
	NER + LongBEL	**70.9** ± 1.7	37.6 ± 1.7	49.2 ± 1.7

Table 4.12: End-to-end named entity recognition (NER) and normalization performance. For each dataset, the *NER only* row reports mention detection performance before normalization. The following rows add a normalization model on top of the same NER predictions. A prediction is considered correct only when both the entity mention and the normalized concept are correctly identified.

4.9 Limitations

Although LongBEL improves document-level consistency in BEL, several limitations remain.

Some limitations discussed in Chapter 3 also apply here. In particular, although we evaluate LongBEL in an end-to-end NER and normalization setting, the experiments are still conducted on public benchmarks rather than on the French discharge summaries studied in Chapter 2. The impact of document-level BEL on these clinical notes therefore remains to be evaluated.

A first limitation specific to LongBEL is related to sequential memory. During inference, each prediction is added to memory and can influence later predictions in the same document. Robust memory training reduces the tendency to blindly copy earlier errors, but it does not remove error propagation completely. If an early prediction is wrong and later mentions are ambiguous, the model may still reuse incorrect information.

A second limitation is inference cost. LongBEL does not add trainable parameters or candidate storage, but it processes longer inputs and updates memory sequentially. In the ensemble setting, several variants are also combined, requiring additional forward passes. As shown in Table 4.11, LongBEL is therefore slower than the corresponding sentence-level SynCABEL models. It is better suited to settings where document-level coherence is important than to applications where maximum throughput is the main constraint.

A third limitation is that the benefit of LongBEL depends on the structure of the documents. LongBEL is most useful when concepts recur, because memory can then provide additional evidence for later mentions. When most concepts appear only once, there is little information to reuse. Moreover, recurrence alone is not sufficient: the analysis on QUAERO-EMEA shows that gains can remain moderate when the absolute number of recurring examples is small or unevenly distributed across semantic groups. LongBEL is therefore expected to be most effective in corpora where concepts recur often enough to provide reliable consistency signals.

This last limitation suggests a promising extension that we do not explore in this chapter: combining SynCABEL and LongBEL. SynCABEL could generate additional examples of recurring concepts, while LongBEL could learn to exploit them for document-level consistency. However, generating synthetic data for LongBEL is more complex than generating sentence-level examples. It would require producing coherent long-context documents with several biomedical mentions, repeated concepts, and consistent annotations across the document. We therefore leave this direction for future work.

5 Conclusion

This chapter introduced **LongBEL**, a document-level generative approach for BEL. Whereas Chapter 3 focused on improving concept coverage with synthetic contextual examples, this chapter addressed another limitation of sentence-level BEL: the lack of consistency across mentions within the same document.

LongBEL combines full-document context with a memory of previous predictions. This allows the model to reuse earlier linking decisions when linking later mentions, especially when the same concept appears several times under different surface forms. To make this memory more robust, we trained the model with out-of-fold predicted memory rather than gold memory, so that the training setting better reflects inference-time conditions.

The experiments show that document-level modeling is useful, but not uniformly so. LongBEL brings the clearest gains on MM-ST21pv, where recurring concepts are frequent and numerous enough to provide stable consistency signals. Gains are smaller or more variable on datasets with fewer recurring examples per semantic group. These results suggest that LongBEL is most useful when consistency is actually needed: concepts must recur within documents, and the training data must contain enough examples for the model to learn how to use these repetitions reliably.

The end-to-end results also show that LongBEL can improve a complete NER and normalization pipeline when document-level signals are strong. However, this comes with higher inference cost: LongBEL does not increase model size or candidate storage, but it processes longer inputs and updates memory sequentially. It is therefore most useful in settings where coherent BEL across a document is more important than maximum throughput.

Together, Chapters 3 and 4 explore two complementary directions for improving biomedical entity linking. SynCABEL addresses the lack of annotated examples and concept coverage, while LongBEL addresses the lack of document-level consistency. The next part concludes the thesis by discussing how these contributions fit into the broader objective of transforming clinical text into reliable structured information.

General Conclusion

1 Answer to the Research Question

The central research question of this thesis was the following: *To what extent can Natural Language Processing (NLP) methods transform clinical notes into structured data suitable for the secondary use of electronic health record (EHR) data?*

The answer provided by this thesis is positive, but conditional. NLP can transform part of the information contained in clinical notes into structured variables, but this transformation remains both partial and imperfect. It does not replace structured EHR data or make the whole content of clinical narratives directly computable. Rather, it extracts selected information from text, with errors that must be taken into account, and makes it usable alongside already structured data.

This idea is illustrated in Figure 5.1, which answers the iceberg representation introduced in Figure 1. Before NLP, clinical notes remain largely below the surface. After NLP, part of this hidden information can be extracted, qualified, and linked to standard concepts. The original notes remain narrative documents, but some of their content becomes structured and can enrich the visible part of the EHR.

Chapter 2 showed that this enrichment is possible in a real-life clinical data warehouse (CDW). Laboratory tests and drug treatments were extracted from French discharge summaries and combined with structured EHR data. The results also showed that text and structured data are complementary: some of the information found in the text was absent from structured tables, while some of the structured information was not found in the text.

The thesis also showed that transforming clinical notes into reusable variables requires a complete pipeline. A clinical mention must first be detected (named entity recognition), interpreted in context (qualification), sometimes associated with a value and/or unit (measurement extraction), and finally linked to a standard concept (biomedical entity linking). Each step can introduce errors, leading to false positives or missed information. Among these steps, biomedical entity linking (BEL) appeared as a major bottleneck: detecting a mention is not sufficient if it cannot be linked to a concept that can be queried and aggregated across patients.

Chapters 3 and 4 therefore focused on improving BEL. SynCABEL used synthetic

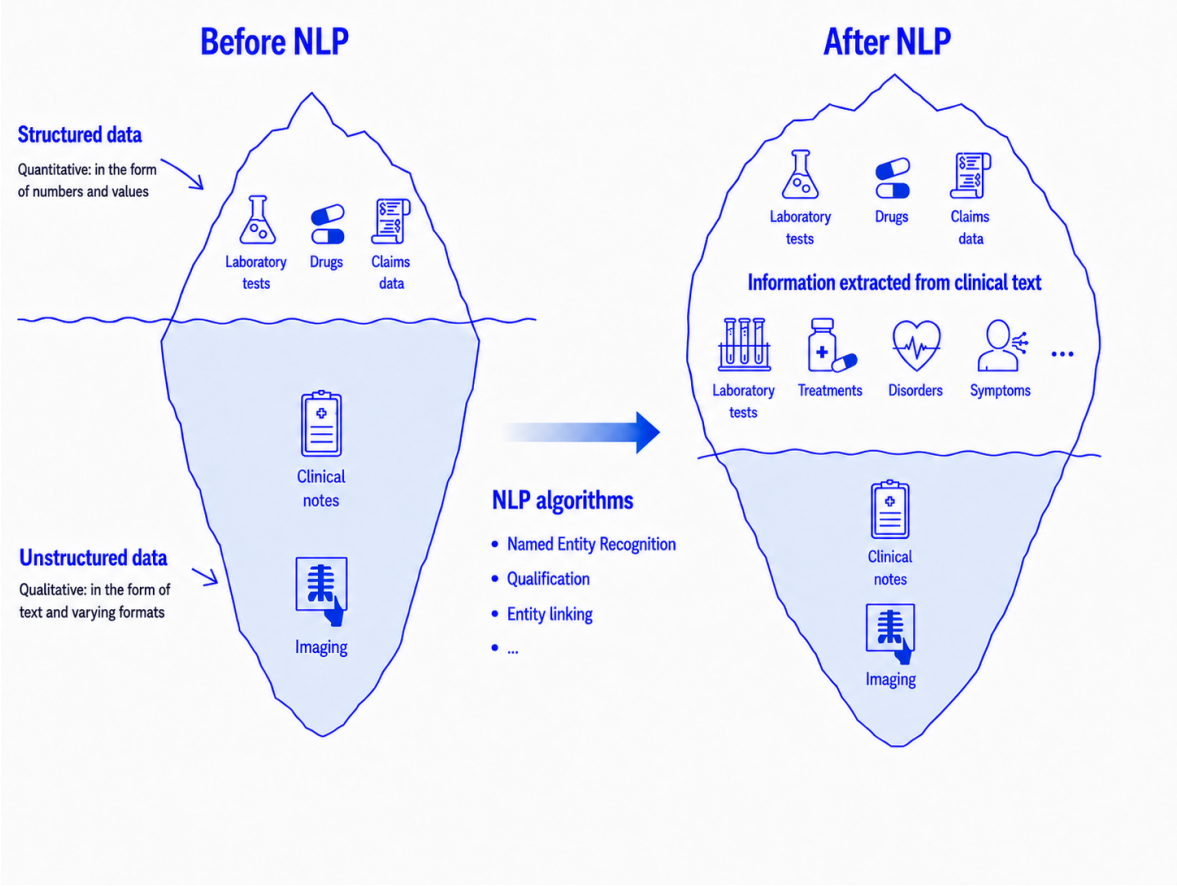


Figure 5.1: Conceptual answer to the iceberg representation introduced in Figure 1. After NLP, part of the information contained in clinical notes can be extracted, qualified, and normalized into structured variables. The original notes remain below the surface, emphasizing that NLP enriches structured EHR data without replacing clinical narratives.

contextualized examples to reduce the need for expert annotations, while LongBEL used full-document context and previous predictions to improve document-level consistency. These approaches improved BEL, but also increased training and inference cost. This trade-off is important in the context of CDWs, where the objective is not to answer isolated questions for a few patients, but to process large numbers of documents at scale.

Finally, the thesis highlights that human annotation remains central. In Chapter 2, supervised named entity recognition (NER) models performed better than rule-based extraction, consistently with previous biomedical NER studies (Chen et al., 2025b; Lee et al., 2020). Annotation is also important beyond NER: qualification requires examples of negation, uncertainty, or temporality in the target clinical context, and previous work on French clinical notes has shown the benefit of machine-learning qualification modules over rule-based alternatives (Petit-Jean et al., 2024). For BEL, this thesis showed that context-aware supervised systems outperform context-free methods. SynCABEL reduces the annotation bottleneck, but it does not remove the need for human annotations: real annotated examples remain necessary to train and evaluate the system.

Overall, NLP can transform part of clinical notes into structured data suitable for

secondary use, provided that the pipeline is computationally realistic for large-scale deployment, supported by appropriate annotations and clinically evaluated.

2 Answer to the Research Objectives

2.1 Objective 1: Transforming Clinical Notes into Structured Variables

The first objective was to evaluate whether clinical notes could be transformed into structured variables in a real EHR setting. This objective was addressed in Chapter 2 through a proof of concept developed in the AP-HP CDW. The pipeline was applied to French discharge summaries and focused on two types of information: laboratory tests and drug treatments.

This work showed that clinically useful information can be extracted from discharge summaries and represented in a structured form, but also that this transformation depends on annotated data. The comparison with the initial rule-based approach showed that existing terminologies alone were not sufficient to obtain the best extraction performance. Manual annotation was therefore needed not only to evaluate the pipeline, but also to define the target information and train supervised models adapted to AP-HP discharge summaries. This also clarifies the scope of the contribution: the pipeline was developed for a specific clinical project, cohort, and document type, rather than a fully unsupervised tool directly generalizable to any clinical variable or document.

The main conclusion is therefore that transforming clinical notes into structured variables is feasible in a real CDW setting, provided that the target information is clearly defined, supported by appropriate annotations, and evaluated for the intended clinical use case.

2.2 Objective 2: Reducing the Annotation Bottleneck in Biomedical Entity Linking

The second objective was to address the need for annotated examples in context-aware BEL. This objective was addressed in Chapter 3 with SynCABEL, a synthetic data augmentation strategy for BEL.

SynCABEL generates contextualized examples from knowledge-base concepts and combines them with human annotations. The goal is to expose the model to a larger part of the target concept space, especially concepts that are absent from the annotated training set. The experiments showed that this strategy improves BEL performance, in particular for unseen concepts, and reduces the amount of human annotation needed to reach strong performance.

The main result of this objective is that synthetic data can make context-aware BEL more scalable. However, it does not remove the need for real annotations. Synthetic examples are useful to increase coverage, but human annotations remain essential.

2.3 Objective 3: Improving Document-Level Consistency in Biomedical Entity Linking

The third objective was to improve the consistency of BEL across complete documents. This objective was addressed in Chapter 4 with LongBEL, a document-level framework

that uses full-document context and a memory of previous predictions.

LongBEL was designed to handle cases where the correct concept cannot always be inferred from the local sentence alone. By reusing earlier linking decisions and broader document context, it helps the model produce more coherent predictions when concepts recur under different surface forms. The experiments showed that these gains are strongest when documents contain repeated concepts, confirming that document-level information is especially useful for consistency.

The main result of this objective is that BEL should not always be treated as a set of independent mention-level decisions. When the document contains useful context or repeated concepts, linking decisions can benefit from being made in relation to one another. This improves coherence, but also introduces additional inference cost, which must be considered for large-scale deployment. The gains obtained with both SynCABEL and LongBEL were demonstrated on public benchmarks and remain to be confirmed on AP-HP discharge summaries.

3 Impact on Clinical Research

Beyond the methodological contributions, an important outcome of this thesis is the reuse of the pipeline studied in Chapter 2 in several clinical research projects. This shows that the work is also a practical tool for helping clinicians and researchers exploit information contained in clinical notes. These projects all rely on the same general idea: transforming selected information from free text into structured variables that can be combined with already structured EHR data.

3.1 Reuse of the Pipeline in Clinical Research Projects

A first example is the published study on diabetic foot ulcers based on the *COhort of DIAbetic Patients* (CODIA) (Julla et al., 2025). CODIA is a project that aims to build a detailed cohort of patients with diabetes using AP-HP EHR data enriched with variables extracted from clinical text. In this first study, the pipeline was applied to discharge summaries from adult diabetic patients hospitalized for a first diabetic foot ulcer. The study included 3,102 patients and reported high 12-month cumulative incidences of death and lower-limb amputation, respectively 21.64% and 24.15%. This illustrates how the pipeline can be reused in another clinical population to enrich patient description with information extracted from clinical text.

A second example is the Immune Mediated Inflammatory Diseases – Identification & Deciphering (IMID-ID) project, led by Pr Raphaèle Seror and Pr Emilie Sbidian. This ongoing project uses the pipeline in a way that is close to Chapter 2: information extracted from discharge summaries is combined with structured data, here from both the AP-HP EHR and the French national health insurance database (*Système National des Données de Santé*). The pipeline is applied at large scale to extract diagnoses, treatments, dosages, dates, and laboratory results from patients with immune-mediated inflammatory diseases. The project also plans to extend the pipeline to new entity types, such as clinical scores, and to evaluate it on project-specific annotations. It therefore illustrates how the pipeline can be reused, adapted, and validated again to build richer large-scale research datasets from clinical text.

Several ongoing projects further illustrate how the pipeline can be adapted to different clinical questions. Two of these projects were conducted by Marine Billon and Léna Levy during research internships following their Master's degrees, while Eliott Laks worked on the AP-HP CDW during a research rotation as part of his medical residency. All three projects were supervised by Christel Gérardin, co-supervisor of this thesis. In the work of Marine Billon on antiphospholipid syndrome, the pipeline is used to improve patient identification and characterization in a disease for which International Classification of Diseases, Tenth Revision (ICD-10) coding lacks specificity. Preliminary results identified a consolidated cohort of 5,949 patients with a certain and non-negated mention of antiphospholipid syndrome and substantially increased the identification of antiphospholipid antibody positivity compared with structured data alone. In the work of Eliott Laks on systemic sclerosis, the pipeline is similarly used to identify patients with scleroderma renal crisis, a rare renal complication without a specific ICD-10 code; no results are available yet. In the work of Léna Levy on antinuclear antibody-positive patients, the pipeline extracts mentions of antinuclear antibodies and systemic autoimmune diseases, filters them using qualification attributes such as certainty and negation, and uses document dates to refine the temporal classification of patients. This led to the reclassification of 26.2% of patients.

Together, these projects show that the pipeline can be reused beyond the initial study to support different clinical research questions. Its main value is to complement structured EHR data with information extracted from clinical text, for example to enrich patient description, improve case identification, or refine patient classification. At the same time, they show that each reuse requires adaptation: the target variables must be defined for the clinical question, and the extracted information must be evaluated in the corresponding setting.

3.2 Potential Uses for Future Clinical Studies

The projects described above represent only a first set of possible applications. A direct future use of the pipeline would be to build enriched research datasets at the scale of the CDW. This would follow the same logic as CODIA or IMID-ID, but on a broader scale: a limited number of variables considered particularly important by the clinical research community would first be selected, precisely defined, annotated, and validated. As in the projects described above, the pipeline would need to be adapted to the target variables, and its use should rely on annotations both for training and evaluation. At the scale of a whole CDW, these annotations should cover a broad range of document types, clinical departments, and patient profiles, in order to make the evaluation more representative of real use. The objective would not be to extract all possible information from the records, but to focus on variables that are important for recurrent clinical research questions and insufficiently captured in structured tables.

A second possible use concerns the identification of similar patients from clinical narratives. Previous work at AP-HP explored this idea by constructing cohorts of similar patients from automatically extracted medical concepts (Gérardin et al., 2022). The general principle is to compare discharge summaries in order to retrieve patients whose clinical situations are close to that of a target patient. During this thesis, we developed a proof-of-concept tool to test whether this approach could be implemented using the information extracted by the pipeline. Figure 5.2 illustrates the prototype inter-

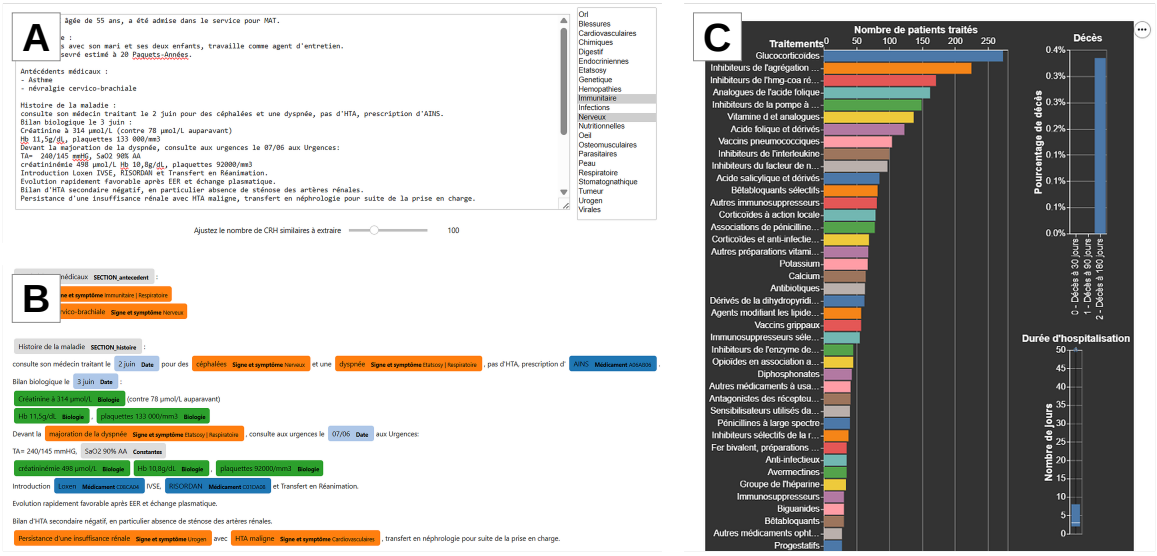


Figure 5.2: Proof-of-concept tool for similar-patient retrieval from discharge summaries. (A) Prototype interface used to input a target discharge summary, select the body systems to compare, and set the number of similar discharge summaries to retrieve. (B) Fictitious example of a target discharge summary after extraction of clinical information by the pipeline. (C) Aggregate summary of treatments and outcomes among fictitious retrieved patients. All patient information shown in this figure is fictitious and used only for illustration. This prototype demonstrates technical feasibility and is now being further developed and evaluated as part of ongoing work.

face. The prototype focused on disorder mentions extracted from discharge summaries. These mentions were represented as embeddings using CODER-all (Yuan et al., 2022c), a model specialized in BEL, and classified into broad body-system categories derived from Medical Subject Headings (MeSH) (Lipscomb, 2000). This classification relied on the classifier developed in the previous similar-patient work from manually annotated data (Gérardin et al., 2022). For each body system, discharge summaries were compared using cosine similarity between disorder embeddings, and Wasserstein distance was used to compare the resulting similarity profiles across body systems. This made it possible to rank discharge summaries according to their similarity to a target patient. An important advantage of this concept-based approach is its explainability: the similarity between two patients can be traced back to specific disorder mentions and their location in the source texts. The resulting retrospective cohort can therefore be inspected and understood rather than being defined only by an opaque global representation.

The tool remains a technical proof of concept, but its development is continuing through the work of Guillaume Faure, whose contribution began during a research internship that I co-supervised with Xavier Tannier and Christel Gérardin and now continues as a PhD project under their supervision, and Anna Nikiforovskaya, a postdoctoral researcher supervised by the same team. This ongoing project will first conduct an intrinsic evaluation in which clinical experts assess the similarity of the retrieved cases to a target patient. An evaluation strategy using a large language model (LLM)-as-a-judge may also be investigated as a complementary evaluation method. A second stage will consist of an extrinsic evaluation in a real clinical setting, for example during multidis-

ciplinary team or internal medicine staff meetings, to determine whether the tool helps clinicians explore comparable cases and support treatment discussions. At this stage, it should therefore be presented as a technical prototype under active development.

A third direction is the emulation of target trials from observational EHR data, an existing methodological approach that retrospectively reproduces the design of a hypothetical randomized clinical trial using routinely collected data. NLP could improve these studies by extracting eligibility criteria, treatment initiation, baseline covariates, contraindications, follow-up information, and outcomes documented only in clinical notes. By providing more complete and clinically precise variables, NLP could help generate stronger preliminary evidence and prioritize the treatments that are most promising to evaluate in prospective randomized trials. It would therefore support treatment selection and trial design, without replacing causal methodology or randomized clinical trials.

A fourth direction is pharmacovigilance and treatment safety research. Clinical notes often describe adverse events, intolerance, treatment interruption, dose changes, and clinical reasoning that are not fully captured in structured data. Extracting these elements could support real-world safety studies, particularly for rare adverse events.

These examples illustrate only some of the possible downstream applications of the pipeline. More broadly, its clinical impact lies in making selected parts of clinical narratives usable at scale. Its value is not to extract all possible information from the EHR, but to support well-defined clinical research questions by producing structured, interpretable, and clinically validated variables from text.

4 Discussion and Perspectives

4.1 Limits of Predefined Structured Extraction

The pipeline studied in this thesis follows a task-specific supervised approach. Before training the model, the information to extract must be defined through annotation guidelines. These guidelines determine which expressions should be marked, which labels should be used, and therefore which variables the model can produce. In Chapter 2, the task focused on laboratory tests and drug treatments because they are clinically important, can be defined relatively clearly, and can usually be mapped to standard concepts in an existing knowledge base (KB). This makes them suitable for large-scale extraction and reuse in a CDW.

Unlike a purely rule-based system, a supervised model can recognize expressions that were not seen exactly during training. However, it can only generalize within the boundaries of the annotation scheme and the distribution represented in the training data. For example, even when disorders are annotated, related expressions such as relapse may fall outside the annotation guidelines or be rare in the training set. The model may therefore fail to detect them, preventing any subsequent normalization. Finally, even when a mention is detected, the target KB may not contain a suitable concept to represent it.

More broadly, the pipeline mainly represents a document as a collection of concepts, possibly enriched with attributes such as negation, temporality, certainty, or measurement values. However, two documents containing the same concepts may describe very

different situations depending on how these concepts are connected and interpreted. It is therefore unrealistic to expect one pipeline to capture all the information contained in a clinical note. Its value lies instead in producing reliable and scalable variables for selected information types defined according to a specific annotation scheme.

4.2 Enriching Structured Variables with Relations

A first extension is to identify relations between extracted concepts. This remains a supervised approach, since the target entities and relations must still be defined and annotated in advance, but it provides a richer representation than a simple list of concepts. A date must be linked to the correct event, a dose to the correct drug, and a symptom to its temporal relation with a disease event. These links are needed to reconstruct clinically meaningful variables from separate mentions.

This direction is also being explored by Guillaume Faure, mentioned above, as part of his ongoing PhD work. His work extends the pipeline presented in Chapter 2 to quantify systemic corticosteroid exposure from French clinical reports. The extraction and normalization components provide a first layer by identifying corticosteroid mentions and dosage-related entities. Relation extraction is then used to associate each corticosteroid mention with its attributes, including dose, route, duration, frequency, and date.

These linked elements can be combined to estimate prednisone-equivalent exposure while accounting for the route of administration, the section of the report, and specific cases such as intravenous corticosteroid pulses. The resulting variable therefore contains more information than the binary observation that a corticosteroid was mentioned. However, relation extraction also introduces new errors: an entity may be linked to the wrong mention, left unlinked, or connected incorrectly across long textual distances, especially for temporal relations. It also remains limited to the entity and relation types defined in advance, so the target information must still be specified according to the downstream clinical question.

4.3 Other Paradigms for Accessing Clinical Information

Relation extraction enriches the structured representation considered in this thesis, but it remains within the same general paradigm: the information of interest is defined in advance and transformed into reusable variables. Many other paradigms can support access to the information contained in clinical narratives, depending on whether the objective is cohort-level analysis, patient-specific exploration, summarization, or clinical decision support. An exhaustive review of these approaches is beyond the scope of this thesis, but one particularly relevant example is currently being developed at AP-HP.

The ChatDPI project uses large language models to facilitate navigation within the EHR of an individual patient. It is not designed as a structured extraction pipeline and does not aim to produce standardized variables across large populations. Instead, it uses the long context capacity of recent models to help clinicians explore a patient record, generate customizable summaries, ask questions in natural language, and trace the answers back to the source documents.

Such a tool is useful when relevant information is spread across many documents or when the question was not defined in advance, for example to reconstruct the history of a complication or understand why a treatment was stopped. ChatDPI and structured

extraction should therefore be viewed as complementary: structured extraction produces validated variables for cohort-level analysis, whereas LLM-based navigation provides more flexible access to the narrative content of one patient record.

4.4 LLM Agents and Structured Extraction as Complementary Approaches

In the coming years, LLM-based agents could become one of the main ways to access information in EHRs and CDWs. Unlike task-specific extraction pipelines, these agents can answer a wide range of clinical questions and extract different types of information without requiring a dedicated model for each variable. They can also work directly with unstructured clinical notes, making it possible to access information that was not anticipated when the system was designed. Their integration into EHR systems is becoming easier thanks to interfaces such as application programming interfaces (APIs) and the Model Context Protocol (MCP), which allow agents to retrieve patient records, query databases, and use external tools. Recent studies have already explored the use of these interfaces to connect LLM agents to EHR systems (Masayoshi et al., 2025; Bellucci et al., 2026).

However, this flexibility comes at a cost. Processing clinical notes with an LLM requires computational resources, and this cost is incurred each time the text is processed. If the same information is needed for several studies or across large patient cohorts, the same documents may have to be processed repeatedly. In contrast, an NLP pipeline can extract and normalize selected information once and store the resulting variables in the CDW. These variables can then be reused through simple database queries, which are much less expensive than repeatedly processing the original text with an LLM. Although the initial extraction also has a cost, it can be shared across many future uses.

For this reason, LLM agents and structured NLP pipelines should not be seen as competing approaches, but as complementary ones. By enriching the CDW with reusable structured variables, NLP pipelines could make LLM agents more efficient. An agent could first query the available structured data and only return to the clinical notes when the information is missing or requires further interpretation. A recent study combining structured and unstructured EHR data with an agentic interface illustrates the potential of this approach (Kim et al., 2026). In this way, structured extraction could reduce the cost of accessing frequently used information, while LLM agents would provide the flexibility needed to answer new clinical questions.

4.5 Perspectives for Biomedical Entity Linking

Chapters 3 and 4 addressed two complementary limitations of biomedical entity linking. SynCABEL increased the coverage of context-aware models by generating synthetic examples for concepts that are poorly represented or absent from human-annotated corpora. LongBEL used full-document context and previous predictions to improve consistency within documents. A natural next step would be to combine these two directions by generating longer synthetic documents containing repeated concepts, abbreviations, synonyms, and mentions that can only be resolved from distant context. Such data could provide supervision for both concept coverage and document-level consistency. Their construction would nevertheless require careful control to ensure that the documents remain coherent and that concept assignments stay consistent across mentions.

However, document-level modeling also comes at a cost. LongBEL increases inference time because each prediction uses longer inputs and may depend on previous predictions in the document. Since many mentions can already be linked correctly from their local context, applying the full model to every mention may not be necessary. A future system could therefore use a fast local model for straightforward cases and activate the more expensive document-level model only for ambiguous, recurring, or uncertain mentions. This decision could rely on confidence scores. Such a design could retain the benefits of LongBEL while reducing its computational cost.

4.6 Overall Perspective

Overall, this thesis supports a balanced view of NLP for clinical notes. NLP can make selected parts of clinical narratives available for secondary use, but it cannot transform the full meaning of a note into structured data. Some information can be represented as standardized variables, some requires relations and temporal modeling, and some, such as clinical reasoning, does not fit naturally into a predefined structured representation.

For large-scale CDW applications, the most realistic goal is therefore not to extract everything, but to extract the information needed for a clearly defined use case. This requires clinicians, data scientists, and NLP researchers to decide which variables matter, how they should be annotated, how errors should be evaluated, and how text-derived variables should be combined with structured EHR data. In this perspective, the thesis provides both an applied demonstration and methodological contributions. The end-to-end pipeline shows how selected information from clinical notes can complement structured EHR data in a real-life CDW. SynCABEL and LongBEL then address two challenges of normalization: limited annotated concept coverage and limited consistency across a document. Together, these contributions support the development of clinical NLP systems that are more scalable, more coherent, and better suited to well-defined secondary uses of EHR data.

Extended French summary – Résumé étendu en français

Le développement des systèmes d'information hospitaliers conduit à une production considérable de données de santé. Ces données constituent aujourd'hui une ressource majeure pour la recherche : elles permettent de mener des études observationnelles, d'améliorer la qualité des soins ou encore de développer des outils d'aide à la décision. Cela permet notamment de mener des études à partir de données déjà disponibles, réduisant ainsi le temps et les coûts nécessaires à leur collecte (Casey et al., 2016). Toutefois, une part importante de l'information contenue dans les dossiers patients informatisés (DPI) reste difficilement exploitable à grande échelle, car elle est exprimée sous forme de texte libre. Cette thèse s'intéresse ainsi à la manière dont le traitement automatique des langues peut contribuer à transformer ces textes cliniques en informations structurées, interprétables et réutilisables.

1 Introduction

1.1 Contexte

1.1.1 Les dossiers patients informatisés

Les DPI regroupent les informations produites au cours de la prise en charge des patients, notamment les données administratives, biologiques, diagnostiques, les comptes rendus textuels et les imageries. Leur principal intérêt pour la recherche réside dans leur collecte à grande échelle, sur de longues périodes et en conditions réelles de soins, permettant notamment d'étudier de larges populations avec des informations cliniques multimodales, y compris chez des patients souvent sous-représentés ou exclus des études cliniques (Callahan et al., 2020).

Cependant, ces données n'ayant pas été initialement recueillies à des fins de recherche, leur réutilisation est méthodologiquement complexe. Elles sont produites au cours de nombreuses activités cliniques, puis extraites et transformées avant d'être intégrées dans des entrepôts de données de santé. À chacune de ces étapes, la disponibilité et la représentation de l'information peuvent varier, ce qui peut notamment entraîner des données manquantes, des complexités temporelles ou des différences entre établissements (Al-Sahab et al., 2024). Ces caractéristiques doivent donc être prises en compte

tout au long de la chaîne de transformation des données, depuis leur production dans les systèmes de soins jusqu'à leur réutilisation au sein d'un entrepôt de données. En France, l'*Entrepôt de Données de Santé* (EDS) de l'Assistance Publique–Hôpitaux de Paris (AP-HP) constitue un exemple de grande infrastructure dédiée à cette réutilisation (Commission nationale de l'informatique et des libertés, 2017).

1.1.2 Données structurées et non structurées

Les données des DPI comprennent des données structurées, telles que les résultats biologiques ou les diagnostics codés, et des données non structurées, notamment les comptes rendus cliniques. Comme l'illustre la Figure 6.1, Les données structurées constituent la partie la plus directement accessible et exploitable des données cliniques, mais elles présentent une double limite : elles ne couvrent qu'une partie de l'information clinique et leur contenu peut être influencé par les pratiques de codage. Certains codes peuvent ainsi être surreprésentés ou, au contraire, sous-représentés en fonction de leur intérêt médico-économique ou des pratiques (Milcent, 2024). Les données structurées peuvent donc fournir une représentation à la fois partielle et biaisée de certaines informations cliniques.

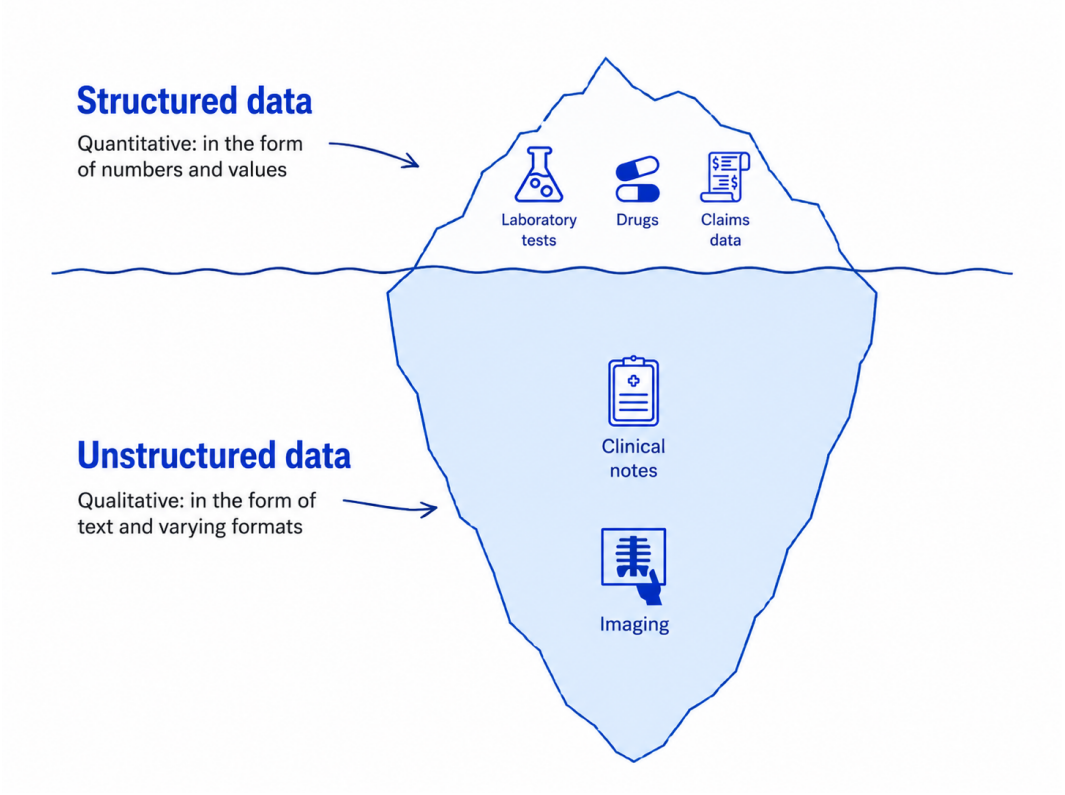


Figure 6.1: Représentation en iceberg des données issues du dossier patient informatisé (DPI) : les données structurées constituent la partie émergée, directement accessible et exploitable, tandis que les données non structurées représentent une part plus riche en informations, mais plus difficile à exploiter.

Les données non structurées peuvent en partie compléter cette information en con-

tenant des éléments qui ne sont pas codés ou difficilement représentables dans des variables prédéfinies, notamment l'histoire médicale, les symptômes, les traitements ou les observations cliniques. Toutefois, leur exploitation à grande échelle nécessite une transformation au moyen de méthodes de traitement automatique des langues (TAL).

1.1.3 Traitement automatique des langues cliniques

Le TAL permet d'exploiter cette information textuelle à grande échelle (Bazoge, 2021). L'extraction d'information vise notamment à identifier les entités cliniques, leurs attributs et les relations qui les relient, puis à normaliser les mentions vers des concepts standardisés. Parmi ces tâches, la reconnaissance d'entités nommées, la qualification contextuelle, l'extraction de relations et la normalisation des entités nommées permettent progressivement de transformer les textes cliniques en informations structurées et réutilisables pour la recherche.

1.2 Problématiques et objectifs

La transformation des notes cliniques en données structurées soulève plusieurs difficultés, cette thèse s'articule ainsi autour de la question suivante :

Dans quelle mesure les méthodes de traitement automatique des langues peuvent-elles transformer les notes cliniques en données structurées adaptées à la réutilisation secondaire des dossiers patients informatisés ?

Afin de répondre à cette question, cette thèse poursuit trois objectifs complémentaires :

1.2.1 Objectif 1 — Évaluer une chaîne de traitement complète pour l'extraction d'informations non structurées en données structurées

Concevoir et évaluer une chaîne complète d'extraction d'information appliquée aux comptes rendus d'hospitalisation, et déterminer dans quelle mesure les informations extraites peuvent compléter les données structurées disponibles dans un entrepôt de données de santé.

1.2.2 Objectif 2 — Réduire le besoin en annotations pour la normalisation

Étudier l'utilisation de données synthétiques contextualisées afin de réduire la dépendance des modèles de normalisation aux annotations expertes et d'améliorer leur couverture de grandes bases de connaissances biomédicales.

1.2.3 Objectif 3 — Améliorer la cohérence de la normalisation à l'échelle d'un document

Étudier dans quelle mesure l'utilisation du contexte du document et des prédictions de normalisation précédentes permettent de produire des annotations plus cohérentes entre les différentes mentions d'un même document.

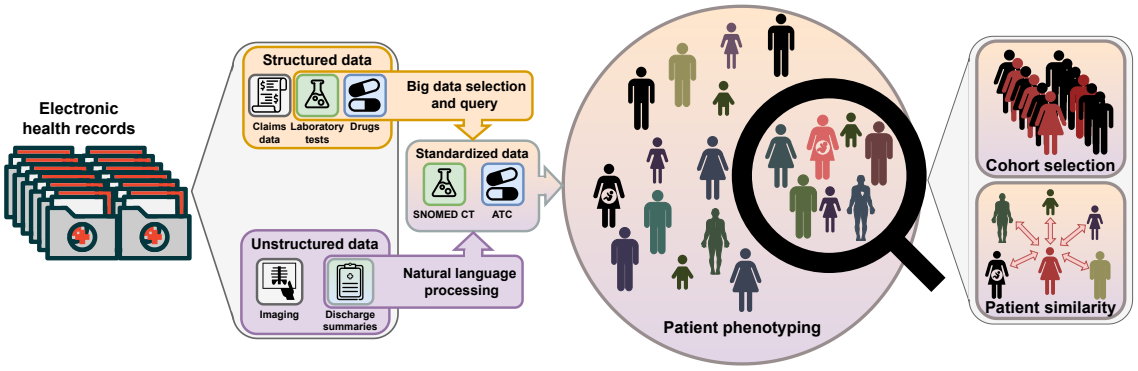


Figure 6.2: Vue d'ensemble de la chaîne de traitement pour l'enrichissement du phénotypage.

2 Une chaîne de TAL pour transformer les comptes rendus en données structurées

Ce premier travail vise à évaluer dans quelle mesure l'extraction d'information à partir de données non structurées permet de compléter les informations déjà disponibles sous forme structurée dans un entrepôt de données de santé, et ainsi d'enrichir le phénotypage des patients. Nous nous intéressons plus particulièrement à l'extraction des examens biologiques et des traitements médicamenteux à partir des comptes rendus d'hospitalisation. Nous présentons d'abord la chaîne de TAL mise en place, avant d'en évaluer les performances et l'apport au phénotypage d'une cohorte de patients atteints de maladies auto-immunes ou inflammatoires systémiques.

2.1 Matériel et méthodes

L'étude a été menée à partir de l'EDS de l'AP-HP et porte sur quatre pathologies : le lupus érythémateux systémique, la sclérodermie systémique, le syndrome des antiphospholipides et l'artérite de Takayasu. La cohorte comprend les DPI de 6 891 patients, regroupant des données structurées, notamment les résultats d'examens biologiques, les prescriptions de médicaments et les codes diagnostiques, ainsi que des données textuelles issues des comptes rendus d'hospitalisation.

Une chaîne complète de TAL a été développée afin d'extraire, de qualifier et de normaliser les informations complémentaires contenues dans les comptes rendus d'hospitalisation, puis de les combiner aux données structurées. Pour les examens biologiques, la chaîne permet d'identifier les mentions correspondant aux examens, d'en extraire et de standardiser les valeurs et les unités, puis de les associer à des concepts standardisés (Figure 6.3). Pour les traitements médicamenteux, elle permet d'identifier les mentions correspondant au nom du médicament, à sa posologie, à sa forme et à son dosage, puis de qualifier leur contexte clinique selon la temporalité (présent, passé ou futur), la certitude (certain, hypothétique ou conditionnel), l'action (début ou arrêt du traitement) et la négation (Figure 6.4). Les mentions de médicaments sont ensuite normalisées selon la classification Anatomical Therapeutic Chemical (ATC), tandis que les examens biologiques sont normalisés vers des concepts du UMLS issus de la terminologie SNOMED CT.

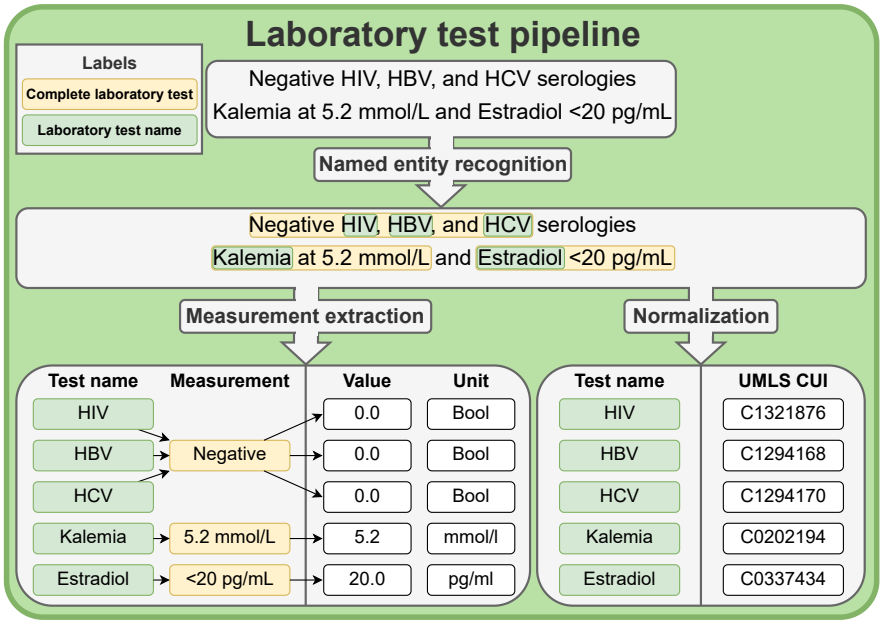


Figure 6.3: Chaîne de traitement des examens biologiques à partir des comptes rendus d'hospitalisation.

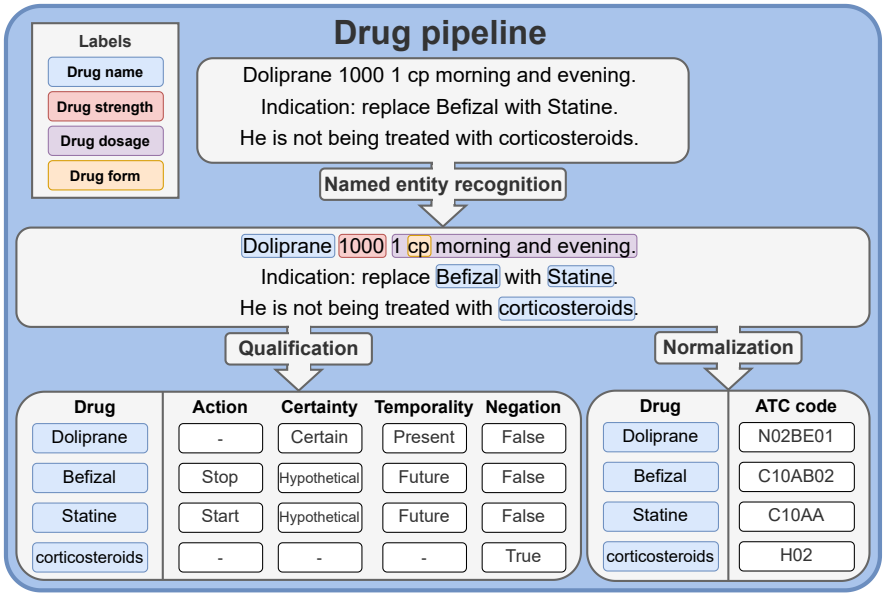


Figure 6.4: Chaîne de traitement des traitements médicamenteux à partir des comptes rendus d'hospitalisation.

La chaîne a d'abord été développée et évaluée sur un sous-ensemble de comptes rendus annotés manuellement par un clinicien, puis appliquée à l'ensemble de la cohorte. L'évaluation a porté à la fois sur les performances des différentes étapes de TAL (reconnaissance d'entités nommées, qualification et normalisation) sur la contribution des données extraites du texte au phénotypage, notamment pour les examens de biologie et les traitements prescrits.

2.2 Résultats et discussion

Les évaluations montrent de bonnes performances pour les différentes tâches de TAL. La reconnaissance d'entités nommées obtient un score F1 global de 88,8%, tandis que la qualification et l'extraction des mesures atteignent respectivement 78,8% et 96,7%. La normalisation est plus performante pour les médicaments que pour les examens biologiques, avec des scores F1 respectifs de 92,9% et 82,2%. Au niveau de la chaîne complète, le score F1 atteint 89,3% pour les médicaments et 71,1% pour les examens biologiques, soulignant l'importance de la normalisation dans la performance de bout en bout.

L'application du système aux 22 194 comptes rendus d'hospitalisation montre que les données textuelles apportent une quantité importante d'informations complémentaires aux données structurées. L'effet est particulièrement marqué pour les auto-anticorps et les traitements. Par exemple, pour le lupus érythémateux systémique, l'ajout des comptes rendus augmente le nombre de patients identifiés comme positifs pour les anticorps antinucléaires de 752 (18,3%) à 2 949 (71,9%), soit 2 197 patients supplémentaires. Pour les anticorps anti-ADN, cela passe de 541 (13,2%) à 2 174 (53,0%), soit 1 633 patients supplémentaires. De même, le nombre de patients traités par hydroxychloroquine passe de 920 (22,4%) à 3 520 (85,8%), soit 2 600 patients supplémentaires. Des gains comparables sont observés pour plusieurs traitements, notamment les glucocorticoïdes et le mycophénolate mofétil, ainsi que pour les trois anticorps étudiés dans la sclérodermie systémique.

L'enrichissement concerne également les valeurs quantitatives des examens de biologie. le nombre de mesures disponibles passe de 31 859 à 40 871 pour la créatinine et de 29 977 à 39 116 pour l'hémoglobine. Les distributions des valeurs extraites sont globalement cohérentes avec celles des données structurées, ce qui montre que les comptes rendus permettent de récupérer des mesures absentes des tables structurées sans modifier sensiblement leur distribution. En revanche, aucun gain n'est observé pour le débit de filtration glomérulaire, dont les mentions n'ont pas pu être correctement normalisées.

Ces résultats montrent donc que les comptes rendus constituent une source pertinente pour compléter les données structurées, mais non pour les remplacer. L'analyse des erreurs souligne en effet que certaines informations présentes dans les données structurées ne sont pas mentionnées dans les comptes rendus, tandis que certaines mentions textuelles restent difficiles à normaliser. L'apport du texte dépend ainsi fortement de la variable considérée et doit être envisagé comme un enrichissement des données structurées plutôt que comme une source exhaustive.

2.3 Conclusion

Ce chapitre montre la faisabilité d'une une chaîne complète de TAL à partir de comptes rendus d'hospitalisation français et son intérêt pour enrichir le phénotypage à partir de données d'un entrepôt de données de santé. La combinaison des données structurées et textuelles permet d'améliorer la couverture de plusieurs variables cliniques importantes. L'étude met toutefois en évidence une difficulté centrale : la normalisation des mentions biomédicales reste un point limitant, en particulier pour les examens biologiques et lorsque l'interprétation dépend du contexte. Cette limite motive le chapitre suivant, consacré à la normalisation des entités biomédicales.

3 Compléter des annotations par des données synthétiques pour la normalisation

Le chapitre précédent a montré l'importance de la normalisation des informations extraites des textes cliniques pour permettre leur réutilisation aux côtés des données structurées. Il a également mis en évidence les limites des approches reposant uniquement sur la mention sans contexte, notamment lorsque celles-ci sont ambiguës. Les méthodes de normalisation exploitant le contexte permettent de lever une partie de ces ambiguïtés, mais nécessitent des exemples annotés. Or, l'annotation doit être faite par des experts, elle est donc coûteuse et ne couvrent généralement qu'une faible proportion des concepts de terminolgies telles que l'UMLS. Ce chapitre introduit ainsi *SynCABEL*, une méthode d'augmentation synthétique visant à améliorer les modèles de normalsation qui utilise le contexte tout en réduisant leur dépendance aux annotations manuelles.

3.1 Matériel et méthodes

SynCABEL s'appuie sur une formulation générative de la normalisation des entités biomédicales, dans laquelle un modèle reçoit une mention accompagnée de son contexte et génère le concept biomédical correspondant. La méthode vise à exploiter l'information contextuelle présente dans les données annotées tout en augmentant le nombre de concepts différents vus par le modèle au cours de son entraînement. Pour cela, elle consiste à générer automatiquement, à partir des informations disponibles dans la terminologie cible, des exemples contextualisés pour des concepts qui ne sont pas nécessairement représentés dans le corpus annoté. La génération de ces exemples repose sur un grand modèle de langage, auquel sont fournies à la fois des informations sur le corpus cible et sur le concept à représenter. Une vue d'ensemble de la méthode est présentée en Figure 6.5.

Plus précisément, le prompt fourni au grand modèle de langage contient des instructions décrivant la tâche de génération, ainsi que plusieurs exemples de phrases issus du corpus annoté, par exemple des résumés d'articles scientifiques biomédicaux en anglais dans le cas du corpus MedMentions (Mohan et al., 2019). À ces éléments s'ajoutent, pour chaque concept pour lequel un exemple doit être généré, les informations disponibles dans la terminologie cible, notamment son nom, ses synonymes, sa définition et son groupe sémantique. À partir de ces informations, le modèle génère une phrase en contexte dans laquelle le concept ciblé est mentionné et associé à son identifiant dans la

terminologie. Ces exemples synthétiques permettent ainsi d'augmenter le nombre de concepts représentés dans les données d'entraînement par rapport aux seuls concepts annotés manuellement, tout en conservant un contexte linguistique cohérent avec le domaine et le corpus cible.

3.2 Matériel et méthodes

SynCABEL s'appuie sur une formulation générative de la normalisation des entités biomédicales, dans laquelle un modèle reçoit une mention accompagnée de son contexte et génère le concept biomédical correspondant. La méthode vise à exploiter l'information contextuelle présente dans les données annotées tout en augmentant le nombre de concepts différents vus par le modèle au cours de son entraînement. Une vue d'ensemble de SynCABEL est présentée en Figure 3.4.

Tout d'abord, la méthode consiste à générer automatiquement, à partir des informations disponibles dans la terminologie cible, des exemples contextualisés pour des concepts qui ne sont pas nécessairement représentés dans le corpus annoté. Plus précisément, un prompt est fourni à un grand modèle de langue. Celui-ci est structuré autour de trois éléments : (i) des instructions décrivant la tâche de génération ; (ii) plusieurs exemples de phrases issus du corpus annoté, par exemple des résumés d'articles scientifiques biomédicaux en anglais dans le cas du corpus MedMentions (Mohan et al., 2019) ; et (iii) les informations disponibles dans la terminologie cible pour le concept considéré, notamment son nom, ses synonymes, sa définition et son groupe sémantique. À partir de ces informations, le modèle génère une phrase en contexte dans laquelle le concept ciblé est mentionné et associé à son identifiant dans la terminologie. Cette procédure permet ainsi de produire des exemples pour des concepts qui ne sont pas ou peu représentés dans les annotations manuelles, tout en conservant un contexte linguistique cohérent avec les données réelles.

Les exemples synthétiques sont ensuite combinés aux données d'entraînement annotées manuellement pour constituer le jeu de données utilisé pour l'entrâinement du modèle de normalisation. Afin de préserver l'importance des données réelles tout en bénéficiant de l'augmentation de couverture apportée par les exemples synthétiques, les deux sources sont utilisées conjointement et les exemples annotés manuellement sont suréchantillonnés de manière à représenter la moitié des instances d'entraînement. Le modèle est ainsi entraîné simultanément sur les exemples issus du corpus et sur les exemples générés.

Lors de l'inférence, la génération libre est contrainte afin d'empêcher le modèle de produire des concepts absents de la terminologie cible. Les concepts disponibles sont d'abord restreints au groupe sémantique considéré. À chaque étape de la génération, le modèle ne peut alors sélectionner que des tokens correspondant à une continuation valide parmi les concepts autorisés. Cette procédure, inspirée de Cao et al. (2021a), garantit ainsi que la séquence générée correspond à un concept valide de la terminologie.

Enfin, afin de compléter l'évaluation fondée sur la correspondance exacte des identifiants, SynCABEL introduit un cadre d'évaluation *LLM-as-a-judge* visant à mesurer la pertinence clinique des prédictions. Cette évaluation repose sur quatre catégories : **Correct**, lorsque le concept prédit est cliniquement approprié, y compris en cas de différence d'identifiant ; **Broad**, lorsqu'il est pertinent mais trop général ; **Narrow**, lorsqu'il est pertinent mais trop spécifique ; et **No relation**, lorsqu'il n'existe pas de relation clinique

pertinente entre les concepts prédit et de référence. Un grand modèle de langue attribue automatiquement ces catégories à partir de définitions précises des classes et d'exemples annotés par des cliniciens. Ce cadre permet ainsi d'évaluer les erreurs de normalisation au-delà de la seule correspondance des codes et de mieux prendre en compte la pertinence clinique des concepts prédits.

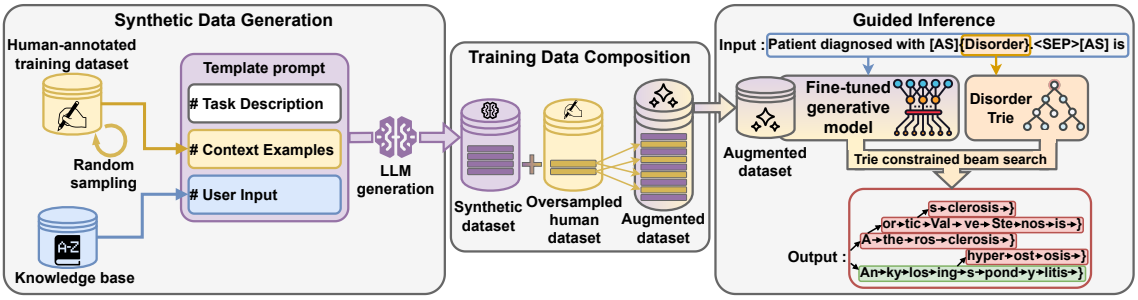


Figure 6.5: Vue d'ensemble de SynCABEL : génération d'exemples contextualisés à partir des concepts d'une base de connaissances, combinaison avec les données annotées et normalisation guidée vers les concepts valides.

3.3 Résultats et discussion

Nous évaluons d'abord l'impact de SynCABEL sur les performances de normalisation d'entités biomédicales. Les performances de SynCABEL sont évaluées sur quatre jeux de données couvrant trois langues et différents types de documents : *MM-ST21pv*, en anglais, constitué de résumés d'articles scientifiques ; *QUAERO-MEDLINE* et *QUAERO-EMEA*, en français, couvrant respectivement des titres d'articles MEDLINE et des documents issus de l'Agence européenne des médicaments ; et *SPACCC*, en espagnol, composé de cas cliniques.

Le Tableau 4.13 montre que SynCABEL améliore de manière systématique les performances des modèles génératifs contextuels sur les quatre jeux de données étudiés. Avec Llama-3-8B, SynCABEL atteint respectivement 75,9% sur MM-ST21pv, 79,7% sur QUAERO-MEDLINE, 79,0% sur QUAERO-EMEA et 67,0% sur SPACCC, soit une moyenne de 75,4%. Il s'agit des meilleures performances obtenues sur les quatre jeux de données. Des améliorations sont également observées avec mBART-large et Llama-3-1B, montrant que le bénéfice de l'augmentation synthétique ne dépend pas exclusivement de l'utilisation d'un modèle de grande taille.

Une analyse selon la présence ou non des concepts dans les données annotées manuellement montre que le bénéfice de SynCABEL provient principalement de sa capacité à améliorer la généralisation vers les concepts absents des annotations manuelles. Les gains sur ces concepts atteignent respectivement +3,7 points sur MM-ST21pv, +5,6 points sur QUAERO-MEDLINE, +12,2 points sur QUAERO-EMEA et +7,0 points sur SPACCC. À l'inverse, les performances sur les concepts déjà observés dans les annotations restent globalement stables. Ces résultats suggèrent ainsi que l'augmentation par données synthétiques permet d'améliorer la généralisation à des concepts absents des annotations manuelles, tout en préservant les performances obtenues sur les concepts déjà représentés dans les données d'entraînement.

Modèle	MM-ST21pv (Anglais)	QUAERO-MEDLINE (Français)	QUAERO-EMEA (Français)	SPACCC (Espagnol)	Moyenne
NORMALISATION SANS CONTEXTE					
SciSpacy	53,8	40,5	37,1	13,2	36,2
SapBERT	65,6	63,2	59,7	33,9	55,6
CODER-all	62,9	72,5	66,9	43,7	61,5
SapBERT-all	64,6	74,7	67,9	47,9	63,8
BERGAMOT	60,9	71,4	63,8	46,0	60,5
NORMALISATION AVEC CONTEXTE					
ArboEL	76,9	71,0	63,0	49,0	65,0
mBART-large	69,6	72,8	69,3	61,8	68,4
+ SynCABEL	71,5	77,1	75,3	64,0	72,0
Llama-1B	73,1	76,2	75,1	63,8	72,1
+ SynCABEL	74,2	77,0	75,8	65,0	73,0
Llama-8B	75,0	77,6	73,8	65,1	72,9
+ SynCABEL	**75,9**	**79,7**	**79,0**	**67,0**	**75,4**

Table 4.13: Performances en Recall@1 des systèmes de normalisation d'entités biomédicales sur les quatre jeux de données évalués. Les lignes « + SynCABEL » correspondent aux modèles entraînés avec augmentation par données synthétiques.

L'étude de l'efficacité des données montre également que SynCABEL peut réduire la quantité d'annotations humaines nécessaire pour atteindre les performances d'un modèle entraîné sur l'ensemble des annotations manuelles. Le modèle augmenté atteint notamment le niveau de référence avec 60% des annotations sur MM-ST21pv et SPACCC, et avec 40% sur QUAERO. Les données synthétiques utilisées seules restent cependant moins performantes que les annotations humaines. Elles constituent donc un complément à la supervision experte plutôt qu'un substitut à celle-ci.

Au-delà de la correspondance exacte des codes, une évaluation clinique plus détaillée montre que de nombreuses prédictions qui ne correspondent pas exactement au code de référence annoté restent néanmoins pertinents. Parmi les 150 erreurs de correspondance exacte analysées par un clinicien, 26,7% ont été jugées correctes, tandis que 28,7% correspondaient à un concept plus général et 21,3% à un concept plus spécifique. Ainsi, 76,7% des prédictions examinées conservaient une relation clinique avec la référence. L'évaluation à plus grande échelle par GPT-5.2 confirme par ailleurs l'intérêt de dépasser la seule correspondance exacte des codes : avec Llama-8B sur SPACCC, l'ajout de SynCABEL augmente la proportion de prédictions jugées *Correct* de 68,9% à 72,6%, tandis que la proportion de prédictions sans relation clinique diminue de 13,8% à 10,5%. Ces résultats suggèrent que l'évaluation par correspondance exacte peut sous-estimer la pertinence clinique réelle des prédictions.

Enfin, l'étude met en évidence un compromis entre performance et coût d'inférence. Les modèles génératifs bénéficient d'une représentation compacte des candidats grâce au décodage contraint, tout en garantissant que les concepts générés appartiennent à la terminologie cible. Les modèles les plus volumineux nécessitent cependant davantage

de mémoire et offrent un débit d'inférence plus faible. Llama-1B présente ainsi un compromis intéressant entre performance et coût de calcul, tandis que Llama-8B obtient les meilleures performances au prix d'une empreinte mémoire plus importante. Le choix du modèle dépend donc du compromis recherché entre performance et contraintes de calcul.

3.4 Conclusion

Ce chapitre introduit SynCABEL, une stratégie d'augmentation par données synthétiques destinée à réduire le besoin en annotations des modèles génératifs de normalisation en contexte. En exploitant les connaissances contenues dans une terminologie biomédicale pour générer des exemples contextualisés, puis en les combinant aux annotations manuelles, la méthode améliore les performances sur plusieurs jeux de données multilingues.

Les expériences montrent également que les données synthétiques peuvent améliorer l'efficacité des annotations humaines, sans toutefois se substituer à l'annotation manuelle experte. Enfin, une évaluation clinique plus détaillée suggère que la qualité de la normalisation ne peut pas être entièrement résumée par la correspondance exacte des codes, certaines prédictions considérées comme incorrectes selon ce critère conservant néanmoins une relation clinique avec le concept de référence.

SynCABEL traite principalement le manque d'annotations manuelles en générant des exemples contextualisés autour de concepts individuels. Les contextes exploités restent cependant locaux, alors que l'interprétation d'une mention peut également dépendre d'informations présentes ailleurs dans le document, notamment de mentions antérieures du même concept ou d'éléments cliniques associés. Le chapitre suivant prolonge donc cette réflexion en s'intéressant à la normalisation à l'échelle du document, en exploitant conjointement le contexte global et les prédictions produites pour les mentions précédentes.

4 LongBEL : normalisation biomédicale cohérente à l'échelle du document

Le chapitre précédent a montré que l'augmentation par données synthétiques permet d'améliorer la performance des modèles de normalisation en contexte. Cependant, SynCABEL reste principalement fondé sur un contexte local et ne tient pas explicitement compte des relations entre les différentes mentions d'un même document. Or, une mention peut être ambiguë lorsqu'elle est considérée isolément, alors que des informations présentes ailleurs dans le document peuvent permettre de l'interpréter correctement. Ce chapitre étudie ainsi l'intérêt d'une modélisation de la normalisation à l'échelle du document et introduit **LongBEL**, une approche générative de la normalisation exploitant conjointement le contexte global du document et les prédictions précédentes du modèle.

4.1 Matériel et méthodes

LongBEL étend la formulation générative introduite dans le chapitre précédent à une prédiction séquentielle à l'échelle du document. Comme illustré en Figure 6.6, les mentions sont traitées dans leur ordre d'apparition et, pour chaque mention, le modèle reçoit

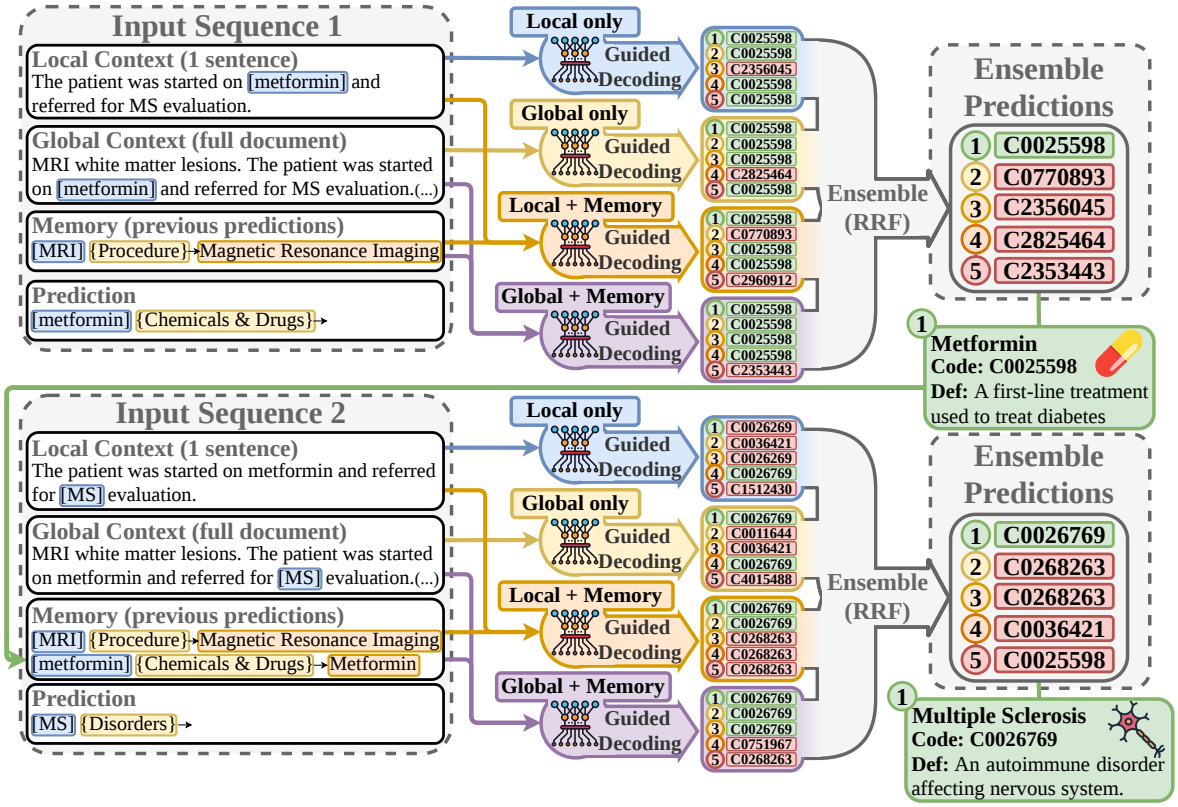


Figure 6.6: Vue d'ensemble de LongBEL : prise en compte du contexte local, du document complet et des prédictions précédentes pour la normalisation séquentielle des mentions, suivie d'une fusion des prédictions par Reciprocal Rank Fusion.

le document complet ainsi qu'une mémoire contenant les mentions précédentes et les concepts qui leur ont été associés. Cette mémoire est mise à jour après chaque prédiction, permettant ainsi de réutiliser les prédictions antérieures lors de la normalisation des mentions suivantes. Les prédictions restent contraintes aux concepts valides de la terminologie au moyen d'un décodage guidé.

Une difficulté importante de cette formulation réside dans le décalage entre l'entraînement et l'inférence : lors de l'inférence, les éléments ajoutés à la mémoire sont les propres prédictions du modèle et peuvent donc être erronés, alors que l'utilisation des concepts de référence pendant l'entraînement produirait une mémoire irréaliste. Pour limiter ce problème, LongBEL construit la mémoire d'entraînement à partir de prédictions. Le jeu d'entraînement est divisé en cinq sous-ensembles ; pour chaque sous-ensemble, un modèle entraîné sur les quatre autres est utilisé pour générer les prédictions correspondantes. Ces prédictions sont ensuite utilisées comme mémoire lors de l'entraînement final de LongBEL, afin de rapprocher les conditions d'entraînement de celles rencontrées à l'inférence.

Enfin, plusieurs variantes du modèle sont combinées afin d'exploiter des signaux complémentaires provenant du contexte local, du contexte global et de la mémoire. Les différentes listes de concepts candidates sont fusionnées à l'aide de la méthode *Reciprocal Rank Fusion* (Cormack et al., 2009).

4.2 Résultats et discussion

LongBEL est évalué sur trois jeux de données multilingues présentant des documents suffisamment longs pour permettre l'étude du contexte documentaire : *MM-ST21pv*, en anglais, constitué de résumés d'articles scientifiques ; *QUAERO-EMEA*, en français, composé de documents issus de l'Agence européenne des médicaments ; et *SPACCC*, en espagnol, constitué de cas cliniques. Contrairement au chapitre précédent, QUAERO-MEDLINE n'est pas retenu en raison de la faible longueur de ses documents, qui ne permet pas d'évaluer de manière pertinente l'apport du contexte global et de la mémoire.

Le Tableau 4.14 présente les principales performances obtenues en Recall@1. Les modèles LongBEL sont comparés aux approches sans contexte, à ArboEL ainsi qu'aux modèles génératifs à contexte local, notamment SynCABEL.

Modèle	MM-ST21pv (Anglais)	QUAERO-EMEA (Français)	SPACCC (Espagnol)	Moyenne
NORMALISATION SANS CONTEXTE				
SciSpacy	53,8	37,1	13,2	34,7
SapBERT	65,6	59,7	33,9	53,1
CODER-all	62,9	66,9	43,7	57,8
SapBERT-all	64,6	67,9	47,9	60,1
BERGAMOT	60,9	63,8	46,0	56,9
NORMALISATION AVEC CONTEXTE				
ArboEL	76,9	63,0	49,0	63,0
mBART-large	69,6	69,3	61,8	66,9
+ SynCABEL	71.5	75.3	64.0	70.3
Llama-1B	73,1	75,1	63,8	70,7
+ SynCABEL	74,2	75,8	65,0	72,4
+ LongBEL	77,6	74,5	63,1	71,7
+ LongBEL Ensemble	79,1	77,2	65,9	73,7
Llama-8B	75,0	73,8	65,1	71,3
+ SynCABEL	75,9	**79,0**	67,0	74,0
+ LongBEL	79,3	75,4	65,3	73,3
+ LongBEL Ensemble	**80,6**	77,6	**67,2**	**74,9**

Table 4.14: Performances en Recall@1 des systèmes de normalisation d'entités biomédicales sur les trois jeux de données évalués. Les lignes en surbrillance correspondent aux modèles LongBEL introduits dans ce chapitre. Les meilleures performances pour chaque jeu de données et pour la moyenne sont indiquées en gras.

Les résultats montrent que LongBEL apporte les gains les plus nets sur MM-ST21pv. Par rapport aux modèles Llama utilisant uniquement le contexte local, le Recall@1 passe de 73,1 à 77,6 pour le modèle 1B et de 75,0 à 79,3 pour le modèle 8B. L'utilisation d'un ensemble de modèles améliore encore ces résultats, avec respectivement 79,1 et 80,6. Sur QUAERO-EMEA et SPACCC, les gains sont plus limités et moins systématiques. Le modèle LongBEL-8B Ensemble obtient néanmoins la meilleure performance moyenne avec 74,9, tandis que SynCABEL reste meilleur sur QUAERO-EMEA et obtient également une performance très proche sur SPACCC.

L'analyse selon la récurrence des concepts apporte une explication à cette dépen-

dance aux jeux de données. Sur MM-ST21pv, où 51,5% des mentions correspondent à des concepts déjà observés dans le document, les gains de LongBEL sont principalement concentrés sur les mentions récurrentes. Pour le modèle 8B, le Recall@1 passe ainsi de 70,1 à 78,4, tandis qu'il reste pratiquement inchangé sur les premières occurrences, de 80,2 à 80,3. Sur QUAERO-EMEA, les mentions récurrentes sont également fréquentes (58,8%), mais leur nombre absolu reste limité et leur répartition entre groupes sémantiques est inégale, ce qui rend les gains moins stables. SPACCC présente quant à lui beaucoup moins de concepts récurrents, ce qui réduit les possibilités d'exploiter la mémoire. Ces résultats montrent que LongBEL agit principalement comme un mécanisme de cohérence à l'échelle du document, son efficacité dépend de la présence de concepts récurrents et du nombre d'exemples disponibles pour permettre au modèle d'apprendre à exploiter ces répétitions.

L'utilisation d'une mémoire construite à partir de prédictions réduit par ailleurs les erreurs en cascade. Cette stratégie diminue le taux de répétition d'une prédiction incorrecte pour les mentions ultérieures d'un même concept, tout en améliorant les performances globales.

Enfin, l'évaluation de bout en bout combinant reconnaissance d'entités et normalisation confirme que LongBEL peut améliorer une chaîne complète lorsque les signaux documentaires sont suffisamment présents. Sur MM-ST21pv, le F1-score passe de 55,4 avec SynCABEL à 58,2 avec LongBEL. En revanche, SynCABEL reste plus performant sur QUAERO-EMEA, avec un F1-score de 64,2 contre 61,5 pour LongBEL, tandis que les deux méthodes obtiennent des performances très proches sur SPACCC. Ces résultats confirment que les deux stratégies répondent à des limitations différentes et doivent davantage être considérées comme complémentaires que comme substituables.

Cette amélioration s'accompagne toutefois d'un coût computationnel accru à l'inférence. LongBEL ne nécessite ni paramètres supplémentaires ni stockage supplémentaire pour les candidats, mais le traitement du document complet et la mise à jour séquentielle de la mémoire réduisent le débit de 69,6 à 48,5 mentions par seconde pour le modèle 1B, et de 38,2 à 15,2 mentions par seconde pour le modèle 8B. Le compromis est donc favorable lorsque la cohérence à l'échelle du document constitue une priorité, mais moins adapté aux applications privilégiant le débit maximal.

4.3 Conclusion

LongBEL montre qu'une approche générative de normalisation peut exploiter efficacement le contexte global et les prédictions antérieures pour améliorer la cohérence des entités au sein d'un même document. Son bénéfice est toutefois fortement dépendant de la récurrence des concepts : lorsque plusieurs mentions d'un même concept sont présentes, les décisions antérieures fournissent une source d'information supplémentaire particulièrement utile.

Ainsi, LongBEL et SynCABEL répondent à deux limitations complémentaires de la normalisation biomédicale. SynCABEL vise à réduire le manque d'annotations grâce à des exemples synthétiques contextualisés, tandis que LongBEL cherche à exploiter l'information distribuée à l'échelle du document pour renforcer la cohérence entre mentions. Cette complémentarité ouvre notamment la voie à une combinaison des deux stratégies, permettant à la fois d'enrichir la couverture des concepts et d'exploiter les récurrences au sein des documents.

5 Synthèse et perspectives

Cette thèse a étudié dans quelle mesure le TAL peut transformer une partie de l'information contenue dans les notes cliniques en données structurées adaptées à la réutilisation secondaire des DPI. Comme l'illustre la Figure 6.7, le TAL permet d'extraire, de qualifier et de normaliser certaines informations textuelles afin de compléter les données déjà structurées du DPI. Cette transformation reste toutefois partielle et imparfaite : toutes les informations contenues dans les notes ne peuvent pas être représentées sous forme de variables structurées. Les méthodes de TAL étudiées dans cette thèse doivent ainsi être considérées comme un moyen d'enrichir les données structurées existantes, sans pour autant permettre de transformer l'ensemble du contenu des notes cliniques en données structurées.

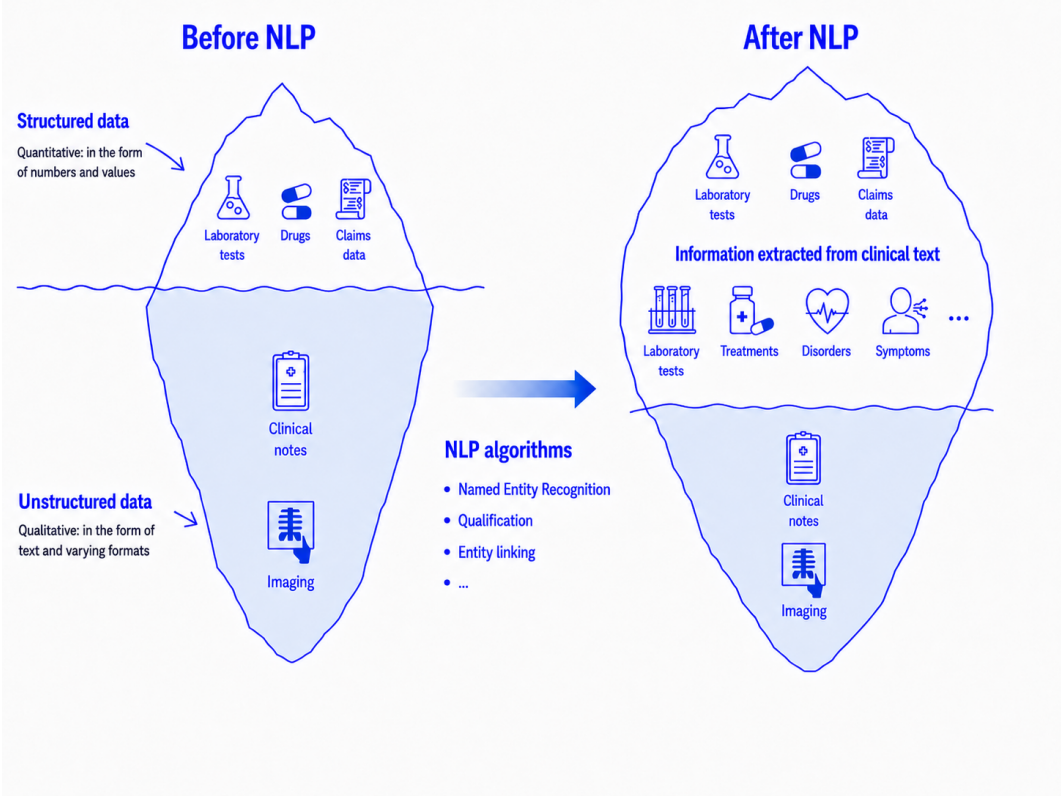


Figure 6.7: De la donnée non structurée à la donnée structurée : le TAL permet d'extraire, qualifier et normaliser une partie des informations contenues dans les notes cliniques.

Le premier volet de la thèse a montré la faisabilité de cette approche dans un entrepôt de données de santé réel. La chaîne de TAL développée permet d'extraire des examens biologiques et des traitements médicamenteux à partir de comptes rendus d'hospitalisation. Ce travail a également mis en évidence le caractère complémentaire des données textuelles et structurées : certaines informations présentes dans les notes ne sont pas disponibles dans les tables structurées, tandis que certaines données structurées ne sont pas retrouvées dans les textes. La chaîne de TAL repose sur plusieurs étapes, allant de la reconnaissance des entités à leur qualification, à l'extraction des valeurs lorsqu'elles existent, puis à leur normalisation. Chaque étape peut introduire

des erreurs et ainsi limiter la qualité de la variable finale. Les travaux ont en particulier montré l'importance de la normalisation, puisque la détection d'une mention ne suffit pas à rendre l'information exploitable à grande échelle si celle-ci ne peut pas être reliée à un concept standardisé.

Les deux chapitres méthodologiques consacrés à la normalisation ont ensuite étudié deux difficultés complémentaires. SynCABEL s'est intéressé à la dépendance des approches contextuelles à la disponibilité d'exemples annotés couvrant suffisamment l'espace des concepts à prédire. L'utilisation d'exemples synthétiques permet d'exposer les modèles à des concepts peu représentés ou absents des données annotées et d'améliorer leurs performances. Cette approche montre ainsi qu'il est possible de réduire la quantité d'annotations humaines nécessaire pour obtenir de bonnes performances, tout en confirmant que les données réelles annotées restent indispensables pour l'entraînement et l'évaluation.

LongBEL a abordé une seconde limitation, liée à la manière dont les mentions sont généralement traitées indépendamment les unes des autres. En exploitant le document complet et les prédictions précédentes du modèle, LongBEL permet de tenir compte d'informations distribuées dans le document et d'améliorer la cohérence entre différentes mentions d'un même concept. Les expériences montrent que ce bénéfice est particulièrement marqué lorsque les concepts sont récurrents dans les documents. Cette approche met ainsi en évidence l'intérêt d'une normalisation globale à l'échelle du document plutôt que locale. En revanche, cette amélioration s'accompagne d'un coût computationnel supplémentaire, qui constitue une contrainte importante dans le contexte d'un traitement à grande échelle des données d'un entrepôt de données de santé.

Pris ensemble, ces travaux montrent que l'amélioration de la normalisation biomédicale repose sur deux leviers distincts. SynCABEL cherche à mieux couvrir l'espace des concepts à prédire malgré des annotations humaines limitées, tandis que LongBEL exploite davantage l'information disponible au sein d'un même document afin de renforcer la cohérence des prédictions. Une perspective naturelle consiste à combiner ces approches, par exemple en générant des documents synthétiques plus longs comportant plusieurs occurrences d'un même concept, des synonymes, des abréviations et des mentions dépendant d'un contexte éloigné. Il serait également intéressant de réduire le coût de LongBEL à l'aide d'une architecture hybride, réservant LongBEL aux mentions ambiguës, récurrentes ou prédite avec un score de confiance faible. D'autres perspectives concernent l'enrichissement des variables extraites par des relations entre entités, afin notamment d'associer dates et événements. À plus long terme, la chaîne complète développée dans cette thèse pourrait contribuer à la construction de cohortes, à la recherche de patients similaires, à l'émulation d'essais cibles ou à la pharmacovigilance.

Ainsi, cette thèse défend une vision pragmatique du TAL appliqué aux notes cliniques. La transformation des textes en données structurées, telle qu'étudiée dans cette thèse, est possible et peut avoir une réelle valeur pour la réutilisation secondaire des données de santé, mais elle nécessite que les types de variables à extraire soient clairement définis avant l'entraînement des algorithmes et que ceux-ci soient évalués dans leur contexte d'utilisation. À terme, l'enjeu n'est donc pas d'extraire tout ce que contient un DPI, mais de développer des méthodes capables d'extraire les informations dont les usagers ont réellement besoin, tout en garantissant que ces informations soient fiables, facilement interprétables et réutilisables.

References

Abas, Ahmed, Ibrahim Elhenawy, Mahinda Zidan, et al. (2021). "BERT-CNN: A Deep Learning Model for Detecting Emotions from Text". In: *Computers, Materials & Continua* 71.2, pp. 2943–2961. DOI: 10.32604/cmc.2022.021671.

Agarwal, Dhruv, Rico Angell, Nicholas Monath, et al. (2022). "Entity Linking via Explicit Mention-Mention Coreference Modeling". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 4644–4658. DOI: 10.18653/v1/2022.naacl-main.343.

Agniel, Denis, Isaac S. Kohane, and Griffin M. Weber (2018). "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study". In: *BMJ* 361, k1479. DOI: 10.1136/bmj.k1479.

Aken, Betty van, Ivana Trajanovska, Amy Siu, et al. (2021). "Assertion Detection in Clinical Notes: Medical Language Models to the Rescue?" In: *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*. Online: Association for Computational Linguistics, pp. 35–40. DOI: 10.18653/v1/2021.nlpmc-1.5.

Aringer, Martin, Karen Costenbader, David Daikh, et al. (2019). "2019 European League Against Rheumatism/American College of Rheumatology Classification Criteria for Systemic Lupus Erythematosus". In: *Arthritis & Rheumatology* 71.9, pp. 1400–1412. DOI: 10.1002/art.40930.

Aronson, Alan R and François-Michel Lang (2010). "An overview of MetaMap: historical perspective and recent advances". In: *Journal of the American Medical Informatics Association* 17.3, pp. 229–236. DOI: 10.1136/jamia.2009.002733.

Bathala, Prasanth, Christophe Ye, Batuhan Nursal, et al. (2025). "BioEL: A Comprehensive Python Package for Biomedical Entity Linking". In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 1709–1721. DOI: 10.18653/v1/2025.findings-naacl.93.

Bazoge, Adrien (2021). "Revue de la littérature : entrepôts de données biomédicales et traitement automatique de la langue (Literature review : biomedical data warehouse and natural language processing)". In: *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 : 23e REncontres jeunes Chercheurs*

en Informatique pour le TAL (RECITAL). Lille, France: ATALA, pp. 96–109. https://aclanthology.org/2021.jeptalnrecital-recital.8/.

Bazoge, Adrien, Emmanuel Morin, Béatrice Daille, et al. (2023). "Applying Natural Language Processing to Textual Data From Clinical Data Warehouses: Systematic Review". In: *JMIR Medical Informatics* 11.1, e42477. DOI: 10.2196/42477.

Bejan, Cosmin Adrian, Fei Xia, Lucy Vanderwende, et al. (2012). "Pneumonia identification using statistical feature selection". In: *Journal of the American Medical Informatics Association* 19.5, pp. 817–823. DOI: 10.1136/amiajnl-2011-000752.

Bellucci, Gianmarco, Wanjun Gu, Peter W. Rose, et al. (2026). "An agentic AI framework connecting language models to electronic health records and a biomedical knowledge graph for real-world evidence". In: *Frontiers in Artificial Intelligence* 9. DOI: 10.3389/frai.2026.1883853.

Bhowmik, Rajarshi, Karl Stratos, and Gerard de Melo (2021). "Fast and Effective Biomedical Entity Linking Using a Dual Encoder". In: *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*. online: Association for Computational Linguistics, pp. 28–37. https://aclanthology.org/2021.louhi-1.4/.

Bodenreider, Olivier (2004). "The Unified Medical Language System (UMLS): integrating biomedical terminology". In: *Nucleic Acids Research* 32.Database issue, pp. D267–270. DOI: 10.1093/nar/gkh061.

Borchert, Florian, Ignacio Llorca, and Matthieu-P. Schapranow (2024). "Improving biomedical entity linking for complex entity mentions with LLM-based text simplification". In: *Database* 2024. DOI: 10.1093/database/baae067.

Brill, Eric (1995). "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging". In: *Computational Linguistics* 21.4, pp. 543–565. https://aclanthology.org/J95-4004/.

Brown, Tom, Benjamin Mann, Nick Ryder, et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Bruijn, Berry de, Colin Cherry, Svetlana Kiritchenko, et al. (2011). "Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010". In: *Journal of the American Medical Informatics Association* 18.5, pp. 557–562. DOI: 10.1136/amiajnl-2011-000150.

Callahan, Alison, Euan Ashley, Somalee Datta, et al. (2023). "The Stanford Medicine data science ecosystem for clinical and translational research". In: *JAMIA Open* 6.3, ooad054. DOI: 10.1093/jamiaopen/ooad054.

Callahan, Alison, Vladimir Polony, José D Posada, et al. (2021). "ACE: the Advanced Cohort Engine for searching longitudinal patient records". In: *Journal of the American Medical Informatics Association* 28.7, pp. 1468–1479. DOI: 10.1093/jamia/ocab027.

Callahan, Alison, Nigam H. Shah, and Jonathan H. Chen (2020). "Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data". In: *Annals of Internal Medicine* 172.11_Supplement, S79–S84. DOI: 10.7326/M19-0873.

Campos, David, Sérgio Matos, and José Luís Oliveira (2012). "Biomedical Named Entity Recognition: A Survey of Machine-Learning Tools". In: *Theory and Applications for Advanced Text Mining*. IntechOpen. DOI: 10.5772/51066.

Cao, Nicola De, Gautier Izacard, Sebastian Riedel, et al. (2021a). *Autoregressive Entity Retrieval*. arXiv. DOI: 10.48550/arXiv.2010.00904.

Cao, Nicola De, Ledell Wu, Kashyap Popat, et al. (2021b). *Multilingual Autoregressive Entity Linking*. arXiv. DOI: 10.48550/arXiv.2103.12528.

Cao, Yixin, Lei Hou, Juanzi Li, et al. (2018). "Neural Collective Entity Linking". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 675–686. https://aclanthology.org/C18-1057/.

Carrino, Casimiro Pio, Joan Llop, Marc Pàmies, et al. (2022). "Pretrained Biomedical Language Models for Clinical NLP in Spanish". In: *Proceedings of the 21st Workshop on Biomedical Language Processing*. Dublin, Ireland: Association for Computational Linguistics, pp. 193–199. DOI: 10.18653/v1/2022.bionlp-1.19.

Carroll, Robert J., Anne E. Eyler, and Joshua C. Denny (2011). "Naïve Electronic Health Record Phenotype Identification for Rheumatoid Arthritis". In: *AMIA Annual Symposium Proceedings* 2011, pp. 189–196. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243261/.

Casey, Joan A., Brian S. Schwartz, Walter F. Stewart, et al. (2016). "Using Electronic Health Records for Population Health Research: A Review of Methods and Applications". In: *Annual Review of Public Health* 37.1, pp. 61–81. DOI: 10.1146/annurev-publhealth-032315-021353.

Chapman, Wendy W., Will Bridewell, Paul Hanbury, et al. (2001). "A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries". In: *Journal of Biomedical Informatics* 34.5, pp. 301–310. DOI: 10.1006/jbin.2001.1029.

Chen, Haihua, Ruochi Li, Ana Cleveland, et al. (2025a). "Enhancing data quality in medical concept normalization through large language models". In: *Journal of Biomedical Informatics* 165, p. 104812. DOI: 10.1016/j.jbi.2025.104812.

Chen, Haihua, Yuhan Zhou, Ruochi Li, et al. (2026). "A comprehensive survey on medical concept normalization: Datasets, techniques, applications, and future directions". In: *Journal of Biomedical Informatics* 178, p. 105005. DOI: 10.1016/j.jbi.2026.105005.

Chen, Qingyu, Yan Hu, Xueqing Peng, et al. (2025b). "Benchmarking large language models for biomedical natural language processing applications and recommendations". In: *Nature Communications* 16.1, p. 3280. DOI: 10.1038/s41467-025-56989-2.

Cocchieri, Alessio, Giacomo Frisoni, Francesco Zangrillo, et al. (2026). "OpenBioNER-v2: A suite of lightweight models for zero-shot medical named entity recognition via type descriptions". In: *Expert Systems with Applications* 318, p. 131725. DOI: 10.1016/j.eswa.2026.131725.

Commission nationale de l'informatique et des libertés (2017). *Délibération n° 2017-013 du 19 janvier 2017 autorisant l'Assistance publique – Hôpitaux de Paris à mettre en œuvre un traitement automatisé de données à caractère personnel ayant pour finalité un entrepôt de données de santé, dénommé « EDS ». (demande d'autorisation n° 1980120)*. https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000034131750.

Commission nationale de l'informatique et des libertés (2021). *Délibération n° 2021-118 du 7 octobre 2021 portant adoption d'un référentiel relatif aux traitements de données à caractère personnel mis en œuvre à des fins de création d'entrepôts de données dans le domaine de la santé*. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000044239566.

Cormack, Gordon V., Charles L A Clarke, and Stefan Buettcher (2009). "Reciprocal rank fusion outperforms condorcet and individual rank learning methods". In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. SIGIR '09. New York, NY, USA. DOI: 10.1145/1571941.1572114.

Daniel, Christel, Nicolas Paris, Olivier Pierre, et al. (2022). "AP-HP Health Data Space (AHDS) to the Test of the Covid-19 Pandemic". In: *Studies in Health Technology and Informatics* 294, pp. 28–32. DOI: 10.3233/SHTI220390.

Daniel, Christel and Elisa Salamanca (2020). "Hospital databases: AP-HP clinical data warehouse". In: *Healthcare and artificial intelligence*. Springer, Cham. DOI: 10.1007/978-3-030-32161-1_8.

Daniel, Christel, Patricia Serre, Nina Orlova, et al. (2019). "Initializing a hospital-wide data quality program. The AP-HP experience". In: *Computer Methods and Programs in Biomedicine* 181, p. 104804. DOI: 10.1016/j.cmpb.2018.10.016.

Dekking, Frederik Michel, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, et al. (2005). "The bootstrap". In: *A Modern Introduction to Probability and Statistics: Understanding Why and How*. London: Springer, pp. 269–284. DOI: 10.1007/1-84628-168-7_18.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

Donnelly, Kevin (2006). "SNOMED-CT: The Advanced Terminology and Coding System for eHealth". In: *Medical and Care Compunetics 3*. IOS Press, pp. 279–290. https://ebooks.iospress.nl/publication/9130.

Doutreligne, Matthieu, Adeline Degremont, Pierre-Alain Jachiet, et al. (2023). "Good practices for clinical data warehouse implementation: A case study in France". In: *PLOS Digital Health* 2.7, e0000298. DOI: 10.1371/journal.pdig.0000298.

Dura, Basile, Charline Jean, Xavier Tannier, et al. (2022). *Learning structures of the French clinical language:development and validation of word embedding models using 21 million clinical reports from electronic health records*. arXiv. DOI: 10.48550/arXiv.2207.12940.

Elkin, Peter L., Sarah Mullin, Jack Mardekian, et al. (2021). "Using Artificial Intelligence With Natural Language Processing to Combine Electronic Health Record's Structured and Free Text Data to Identify Nonvalvular Atrial Fibrillation to Decrease Strokes and Death: Evaluation and Case-Control Study". In: *Journal of Medical Internet Research* 23.11, e28946. DOI: 10.2196/28946.

European Parliament and Council of the European Union (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free move-*

ment of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Eviatar, Tali, Roni Yahalom, Idit Livnat, et al. (2024). "Real-world treatment patterns in patients with systemic lupus erythematosus: associations with comorbidities and damage." In: *Lupus science & medicine* 11.2. DOI: 10.1136/lupus-2024-001266.

Fanouriakis, Antonis, Myrto Kostopoulou, Alessia Alunno, et al. (2019). "2019 update of the EULAR recommendations for the management of systemic lupus erythematosus". In: *Annals of the Rheumatic Diseases* 78.6, pp. 736–745. DOI: 10.1136/annrheumdis-2019-215089.

Fellegi, Ivan P. and Alan B. Sunter (1969). "A Theory for Record Linkage". In: *Journal of the American Statistical Association* 64.328, pp. 1183–1210. DOI: 10.1080/01621459.1969.10501049.

Fraile Navarro, David, Kiran Ijaz, Dana Rezazadegan, et al. (2023). "Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review". In: *International Journal of Medical Informatics* 177, p. 105122. DOI: 10.1016/j.ijmedinf.2023.105122.

Frankovich Jennifer, Longhurst Christopher A., and Sutherland Scott M. (2011). "Evidence-Based Medicine in the EMR Era". In: *New England Journal of Medicine* 365.19, pp. 1758–1759. DOI: 10.1056/NEJMp1108726.

French, Evan and Bridget T. McInnes (2023). "An overview of biomedical entity linking throughout the years". In: *Journal of Biomedical Informatics* 137, p. 104252. DOI: 10.1016/j.jbi.2022.104252.

Friedman, Carol, Lyudmila Shagina, Socrates A. Socratous, et al. (1996). "A WEB-Based Version of MedLEE: A Medical Language Extraction and Encoding System". In: *Proceedings of the AMIA Annual Fall Symposium*, p. 938. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233000/.

Ganea, Octavian-Eugen and Thomas Hofmann (2017). "Deep Joint Entity Disambiguation with Local Neural Attention". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2619–2629. DOI: 10.18653/v1/D17-1277.

Garcelon, Nicolas, Antoine Neuraz, Vincent Benoit, et al. (2017). "Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse". In: *Journal of the American Medical Informatics Association* 24.3, pp. 607–613. DOI: 10.1093/jamia/ocw144.

Ge, Yao, Yuting Guo, Sudeshna Das, et al. (2026). "Improving few-shot named entity recognition for large language models using structured dynamic prompting with retrieval augmented generation". In: *npj Artificial Intelligence* 2.1, p. 39. DOI: 10.1038/s44387-025-00062-2.

Gérardin, Christel, Arthur Mageau, Arsène Mékinian, et al. (2022). "Construction of Cohorts of Similar Patients From Automatic Extraction of Medical Concepts: Phenotype Extraction Study". In: *JMIR medical informatics* 10.12, e42379. DOI: 10.2196/42379.

Gerner, Martin, Goran Nenadic, and Casey M. Bergman (2010). "LINNAEUS: a species name identification system for biomedical literature". In: *BMC bioinformatics* 11, p. 85. DOI: 10.1186/1471-2105-11-85.

Gombar, Saurabh, Alison Callahan, Robert Califf, et al. (2019). "It is time to learn from patients like mine". In: *npj Digital Medicine* 2.1, pp. 1–3. DOI: 10.1038/s41746-019-0091-3.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, et al. (2024). *The Llama 3 Herd of Models*. arXiv. DOI: 10.48550/arXiv.2407.21783.

Gretton, Arthur, Karsten M. Borgwardt, Malte J. Rasch, et al. (2012). "A Kernel Two-Sample Test". In: *Journal of Machine Learning Research* 13.25, pp. 723–773. http://jmlr.org/papers/v13/gretton12a.html.

Gu, Yu, Robert Tinn, Hao Cheng, et al. (2021). "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing". In: *ACM Transactions on Computing for Healthcare (HEALTH)* 3.1, 2:1–2:23. DOI: 10.1145/3458754.

Hanisch, Daniel, Katrin Fundel, Heinz-Theodor Mevissen, et al. (2005). "ProMiner: rule-based protein and gene entity recognition". In: *BMC Bioinformatics* 6.Suppl 1, S14. DOI: 10.1186/1471-2105-6-S1-S14.

Harkema, Henk, John N. Dowling, Tyler Thornblade, et al. (2009). "ConText: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports". In: *Journal of Biomedical Informatics* 42.5, pp. 839–851. DOI: 10.1016/j.jbi.2009.05.002.

Harron, Katie, Angie Wade, Ruth Gilbert, et al. (2014). "Evaluating bias due to data linkage error in electronic healthcare records". In: *BMC medical research methodology* 14, p. 36. DOI: 10.1186/1471-2288-14-36.

Haute Autorité de Santé (2020). *Artérite de Takayasu*. https://www.has-sante.fr/jcms/p_3148994/fr/arterite-de-takayasu.

Haute Autorité de Santé, H.A.S. (2018). *Sclérodermie Systémique*. Guide maladie chronique. Saint-Denis La Plaine: Haute Autorité de Santé, H.A.S. https://www.has-sante.fr/jcms/c_717292/fr/sclerodermie-systemique.

Haute Autorité de Santé, H.A.S. (2022). *Syndrome des Anti-Phospholipides de l'adulte et de l'enfant*. Guide maladie chronique. Saint-Denis La Plaine: Haute Autorité de Santé, H.A.S. https://www.has-sante.fr/jcms/p_3375791/fr/syndrome-des-anti-phospholipides-de-l-adulte-et-de-l-enfant.

Haute Autorité de Santé, H.A.S. (2024). *Lupus Systémique de l'adulte et de l'enfant*. Guide maladie chronique. Saint-Denis La Plaine: Haute Autorité de Santé, H.A.S. https://www.has-sante.fr/jcms/p_3493410/fr/lupus-systemique-de-l-adulte-et-de-l-enfant.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao, et al. (2022). "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing". In: *International Conference on Learning Representations*. https://openreview.net/forum?id=sE7-XhLxHA.

Henke, Elisa, Michele Zoch, Yuan Peng, et al. (2024). "Conceptual design of a generic data harmonization process for OMOP common data model". In: *BMC medical informatics and decision making* 24.1, p. 58. DOI: 10.1186/s12911-024-02458-7.

Honnibal, Matthew, Ines Montani, Matthew Honnibal, et al. (2018). *explosion/spaCy: v2.0.11: Alpha Vietnamese support, fixes to vectors, improved errors and more*. Zenodo. DOI: 10.5281/zenodo.1212304.

Hripcsak, George, Jon D Duke, Nigam H Shah, et al. (2015). "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers".

In: *Studies in health technology and informatics* 216, pp. 574–578. https://pmc.ncbi.nlm.nih.gov/articles/PMC4815923/.

Hu, Yan, Xu Zuo, Yujia Zhou, et al. (2026). "Information extraction from clinical notes: are we ready to switch to large language models?" In: *Journal of the American Medical Informatics Association* 33.3, pp. 553–562. DOI: 10.1093/jamia/ocaf213.

Huang, Kexin, Jaan Altosaar, and Rajesh Ranganath (2019). "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission". In: *arXiv:1904.05342*.

Huang, Zhiheng, Wei Xu, and Kai Yu (2015). *Bidirectional LSTM-CRF Models for Sequence Tagging*. arXiv. DOI: 10.48550/arXiv.1508.01991.

Idnay, Betina, Gongbo Zhang, Fangyi Chen, et al. (2025). "Mini-mental status examination phenotyping for Alzheimer's disease patients using both structured and narrative electronic health record features". In: *Journal of the American Medical Informatics Association* 32.1, pp. 119–128. DOI: 10.1093/jamia/ocae274.

International Organization for Standardization (2005). *Health informatics — Electronic health record — Definition, scope and context*. ISO Technical Report ISO/TR 20514:2005. International Organization for Standardization. https://www.iso.org/obp/ui/#iso:std:iso:tr:20514:ed-1:v1:en.

Jiang, Min, Yonghui Wu, Anushi Shah, et al. (2014). "Extracting and standardizing medication information in clinical text – the MedEx-UIMA system". In: *AMIA Summits on Translational Science Proceedings* 2014, pp. 37–42. https://pmc.ncbi.nlm.nih.gov/articles/PMC4419757/.

Julla, Jean-Baptiste, Théo Jolivet, Candice Estellat, et al. (2025). "Incidence of death and amputation in patients with a first diabetic foot ulcer: results from the CODIA cohort". In: *Diabetes & Metabolism* 51.6, p. 101700. DOI: 10.1016/j.diabet.2025.101700.

Kang, Yanna Shen and Mehmet Kayaalp (2013). "Extracting laboratory test information from biomedical text". In: *Journal of Pathology Informatics* 4.1, p. 23. DOI: 10.4103/2153-3539.117450.

Keloth, Vipina K, Yan Hu, Qianqian Xie, et al. (2024). "Advancing entity recognition in biomedicine via instruction tuning of large language models". In: *Bioinformatics* 40.4, btae163. DOI: 10.1093/bioinformatics/btae163.

Khurshid, Shaan, Christopher Reeder, Lia X. Harrington, et al. (2022). "Cohort design and natural language processing to reduce bias in electronic health records research". In: *npj Digital Medicine* 5.1, pp. 1–14. DOI: 10.1038/s41746-022-00590-0.

Kim, Chanhwi, Hyunjae Kim, Sihyeon Park, et al. (2025). "Learning from Negative Samples in Biomedical Generative Entity Linking". In: *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria: Association for Computational Linguistics, pp. 10714–10730. https://aclanthology.org/2025.findings-acl.558/.

Kim, Edward, Richard Foty, Avnesh S. Thakor, et al. (2026). "Computable longitudinal patient journeys from structured and unstructured EHR data". In: *Nature Medicine*, pp. 1–11. DOI: 10.1038/s41591-026-04695-x.

Kim, Youngjun, Ellen Riloff, and Stéphane Meystre (2011). "Improving Classification of Medical Assertions in Clinical Notes". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 311–316. https://aclanthology.org/P11-2054/.

Labrak, Yanis, Adrien Bazoge, Richard Dufour, et al. (2023). "DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 16207–16221. DOI: 10.18653/v1/2023.acl-long.896.

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 282–289. http://www.cs.columbia.edu/~jebara/6772/papers/crf.pdf.

Laforet, Jeremy (2017). *BM04 – Échanges et stockage de données*. https://www.utc.fr/~jlaforet/cours_BM04/cours_2.html#/3/4.

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, et al. (2016). "Neural Architectures for Named Entity Recognition". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 260–270. DOI: 10.18653/v1/N16-1030.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, et al. (2020). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4, pp. 1234–1240. DOI: 10.1093/bioinformatics/btz682.

Levenshtein, Vladimir I. (1966). "Binary Codes Capable of Correcting Deletions, Insertions and Reversals". In: *Soviet Physics Doklady* 10.8, pp. 707–710. https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf.

Lewis, Mike, Yinhan Liu, Naman Goyal, et al. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. arXiv. DOI: 10.48550/arXiv.1910.13461.

Liao, Katherine P., Tianxi Cai, Vivian Gainer, et al. (2010). "Electronic medical records for discovery research in rheumatoid arthritis". In: *Arthritis Care & Research* 62.8, pp. 1120–1127. DOI: 10.1002/acr.20184.

Lima-López, Salvador, Eulalia Farré-Maduell, Luis Gasco Sanchez, et al. (2023a). "Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text". In: *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the Era of Generative Models*. DOI: 10.5281/zenodo.10104547.

Lima-López, Salvador, Eulàlia Farré-Maduell, Luis Gasco, et al. (2023b). "Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023". In: *Working Notes of CLEF*. https://ceur-ws.org/Vol-3497/paper-002.pdf.

Lin, Zhenxi, Ziheng Zhang, Xian Wu, et al. (2024). "Biomedical Entity Linking as Multiple Choice Question Answering". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, pp. 2390–2396. https://aclanthology.org/2024.lrec-main.214/.

Lipscomb, Carolyn E. (2000). "Medical Subject Headings (MeSH)". In: *Bulletin of the Medical Library Association* 88.3, pp. 265–266. https://pmc.ncbi.nlm.nih.gov/articles/PMC35238/.

Liu, Fangyu, Ehsan Shareghi, Zaiqiao Meng, et al. (2021a). “Self-Alignment Pretraining for Biomedical Entity Representations”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 4228–4238. DOI: 10.18653/v1/2021.naacl-main.334.

Liu, Fangyu, Ivan Vulić, Anna Korhonen, et al. (2021b). “Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, pp. 565–574. DOI: 10.18653/v1/2021.acl-short.72.

Liu, Sijia, Liwei Wang, Donna Ihrke, et al. (2017). “Correlating Lab Test Results in Clinical Notes with Structured Lab Data: A Case Study in HbA1c and Glucose”. In: *AMIA Summits on Translational Science Proceedings* 2017, pp. 221–228. https://pmc.ncbi.nlm.nih.gov/articles/PMC5543347/.

Lowe, Henry J., Todd A. Ferris, Penni M. Hernandez, et al. (2009). “STRIDE – An Integrated Standards-Based Translational Research Informatics Platform”. In: *AMIA Annual Symposium Proceedings*. San Francisco, CA, USA, pp. 391–395. https://pmc.ncbi.nlm.nih.gov/articles/PMC2815452/.

Ma, Xuezhe and Eduard Hovy (2016). “End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1064–1074. DOI: 10.18653/v1/P16-1101.

Mahajan, Diwakar, Jennifer J. Liang, and Ching-Huei Tsou (2022). “Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives”. In: *AMIA Annual Symposium Proceedings* 2021, pp. 833–842. https://pmc.ncbi.nlm.nih.gov/articles/PMC8861744/.

Mahajan, Diwakar, Jennifer J. Liang, Ching-Huei Tsou, et al. (2023). “Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes”. In: *Journal of Biomedical Informatics* 144, p. 104432. DOI: 10.1016/j.jbi.2023.104432.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, et al. (2020). “CamemBERT: a Tasty French Language Model”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7203–7219. DOI: 10.18653/v1/2020.acl-main.645.

Masayoshi, Kanato, Masahiro Hashimoto, Ryoichi Yokoyama, et al. (2025). *EHR-MCP: Real-world Evaluation of Clinical Information Retrieval by Large Language Models via Model Context Protocol*. arXiv. https://arxiv.org/abs/2509.15957.

Mikolov, Tomas, Kai Chen, Greg Corrado, et al. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv. DOI: 10.48550/arXiv.1301.3781.

Milcent, Carine (2024). “Bias due to re-used databases: Coding in hospital for extremely vulnerable patients”. In: *Health Policy and Technology* 13.2, p. 100851. DOI: 10.1016/j.hlpt.2024.100851.

Ministère du Travail, de la Santé et des Solidarités and Agence du Numérique en Santé (2025). *Catalogue des terminologies*. https://smt.esante.gouv.fr/catalogue-des-terminologies/.

Miranda-Escalada, Antonio, Luis Gasco, Salvador Lima-López, et al. (2022). "Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources". In: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*. https://ceur-ws.org/Vol-3180/paper-11.pdf.

Mohan, Sunil and Donghui Li (2019). *MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts*. arXiv. DOI: 10.48550/arXiv.1902.09476.

Moqurrab, Syed Atif, Umair Ayub, Adeel Anjum, et al. (2021). "An Accurate Deep Learning Model for Clinical Entity Recognition From Clinical Notes". In: *IEEE Journal of Biomedical and Health Informatics* 25.10, pp. 3804–3811. DOI: 10.1109/JBHI.2021.3099755.

Mouthon, Luc, Guillaume Bussone, Alice Berezné, et al. (2014). "Scleroderma renal crisis". In: *The Journal of Rheumatology* 41.6, pp. 1040–1048. DOI: 10.3899/jrheum.131210.

Murphy, Shawn N., Griffin Weber, Michael Mendis, et al. (2010). "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)". In: *Journal of the American Medical Informatics Association* 17.2, pp. 124–130. DOI: 10.1136/jamia.2009.000893.

Nadeau, David and Satoshi Sekine (2007). "A survey of named entity recognition and classification". In: *Lingvisticæ Investigationes* 30.1, pp. 3–26. DOI: 10.1075/li.30.1.03nad.

Naguib, Marco, Xavier Tannier, and Aurélie Névéol (2024). "Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, pp. 6829–6852. DOI: 10.18653/v1/2024.findings-emnlp.400.

Neumann, Mark, Daniel King, Iz Beltagy, et al. (2019). "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing". In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, pp. 319–327. DOI: 10.18653/v1/W19-5034.

Névéol, Aurélie, Hercules Dalianis, Sumithra Velupillai, et al. (2018). "Clinical Natural Language Processing in languages other than English: opportunities and challenges". In: *Journal of Biomedical Semantics* 9.1, p. 12. DOI: 10.1186/s13326-018-0179-8.

Névéol, Aurélie, Cyril Grouin, Jeremy Leixa, et al. (2014). "The QUAERO French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization". In: *Proc of BioTextMining Work*, pp. 24–30.

Nguyen, Duc Hau (2024). "Making AI Understandable for Humans : the Plausibility of Attention-Based Explanations in Natural Language Processing". Theses. INSA de Rennes. DOI: 10.70675/97b1b150zc009z4f96z9ec6z2790870387a0.

Papez, Vaclav, Maxim Moinat, Stefan Payralbe, et al. (2021). "Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP common data model: a case study in heart failure". In: *JAMIA Open* 4.3, ooab001. DOI: 10.1093/jamiaopen/ooab001.

Park, Yesol, Gyujin Son, and Mina Rho (2024). "Biomedical Flat and Nested Named Entity Recognition: Methods, Challenges, and Advances". In: *Applied Sciences* 14.20, p. 9302. DOI: 10.3390/app14209302.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

Petit-Jean, Thomas, Christel Gérardin, Emmanuelle Berthelot, et al. (2024). "Collaborative and privacy-enhancing workflows on a clinical data warehouse: an example developing natural language processing pipelines to detect medical conditions". In: *Journal of the American Medical Informatics Association* 31.6, pp. 1280–1290. DOI: 10.1093/jamia/ocae069.

Petit-Jean, Thomas, Adam Remaki, Vincent Maladière, et al. (2022). *eds-scikit: data analysis on OMOP databases*. DOI: 10.5281/zenodo.7401549.

Priou, Sonia, Emmanuelle Kempf, Rémi Flicoteaux, et al. (2024). "'*Where have my patients gone?*': A simulation study on real-world data processing in Clinical Data Warehouses". In: *Health Policy and Technology* 13.3, p. 100893. DOI: 10.1016/j.hlpt.2024.100893.

Priou, Sonia, Guillaume Lame, Marija Jankovic, et al. (2023). "Why Are Data Missing in Clinical Data Warehouses? A Simulation Study of How Data Are Processed (and Can Be Lost)". In: *Studies in Health Technology and Informatics* 302, pp. 202–206. DOI: 10.3233/SHTI230103.

Ramshaw, Lance and Mitch Marcus (1995). "Text Chunking using Transformation-Based Learning". In: *Third Workshop on Very Large Corpora*. https://aclanthology.org/W95-0107/.

Ratinov, Lev and Dan Roth (2009). "Design Challenges and Misconceptions in Named Entity Recognition". In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*. Boulder, Colorado: Association for Computational Linguistics, pp. 147–155. https://aclanthology.org/W09-1119.

Ratinov, Lev, Dan Roth, Doug Downey, et al. (2011). "Local and Global Algorithms for Disambiguation to Wikipedia". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 1375–1384. https://aclanthology.org/P11-1138/.

Reich, Christian, Anna Ostropolets, Patrick Ryan, et al. (2024). "OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization". In: *Journal of the American Medical Informatics Association* 31.3, pp. 583–590. DOI: 10.1093/jamia/ocad247.

Remaki, Adam, Christel Gérardin, Eulàlia Farré-Maduell, et al. (2026a). "SynCABEL: Synthetic Contextualized Augmentation for Biomedical Entity Linking". In: *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*. Ed. by James Kwok. International Joint Conferences on Artificial Intelligence Organization, pp. 6859–6867. DOI: 10.24963/ijcai.2026/763.

Remaki, Adam, Benoît Playe, Paul Bernard, et al. (2023a). *Adjusting for the progressive digitization of health records: working examples on a multi-hospital clinical data warehouse*. medRxiv. DOI: 10.1101/2023.08.17.23294220.

Remaki, Adam, Xavier Tannier, and Christel Gérardin (2026b). *LongBEL: Long-Context and Document-Consistent Biomedical Entity Linking*. arXiv. DOI: 10.48550/arXiv.2605.13451.

Remaki, Adam, Jacques Ung, Pierre Pages, et al. (2025). “Improving Phenotyping of Patients With Immune-Mediated Inflammatory Diseases Through Automated Processing of Discharge Summaries: Multicenter Cohort Study”. In: *JMIR Medical Informatics* 13.1, e68704. DOI: 10.2196/68704.

Remaki, Adam, Simon Vittoz, and Charline Jean (2023b). *aphp/edsteva*. Zenodo. DOI: 10.5281/zenodo.10074576.

Richesson, Rachel L, W Ed Hammond, Meredith Nahm, et al. (2013). “Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory”. In: *Journal of the American Medical Informatics Association* 20.e2, e226–e231. DOI: 10.1136/amiajnl-2013-001926.

Ruas, Pedro, Sofia Conceição, and Francisco M. Couto (2026). “Systematic Review of Named Entity Linking and Knowledge Organisation Systems in Biomedical and Clinical Domains”. In: *ACM Computing Surveys* 58.10, 259:1–259:36. DOI: 10.1145/3796222.

Al-Sahab, Ban, Alan Leviton, Tobias Loddenkemper, et al. (2024). “Biases in Electronic Health Records Data for Generating Real-World Evidence: An Overview”. In: *Journal of Healthcare Informatics Research* 8.1, pp. 121–139. DOI: 10.1007/s41666-023-00153-2.

Sakhovskiy, Andrey, Natalia Semenova, Artur Kadurin, et al. (2024). “Biomedical Entity Representation with Graph-Augmented Multi-Objective Transformer”. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Mexico City, Mexico: Association for Computational Linguistics, pp. 4626–4643. DOI: 10.18653/v1/2024.findings-naacl.288.

Savova, Guergana K, James J Masanz, Philip V Ogren, et al. (2010). “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications”. In: *Journal of the American Medical Informatics Association* 17.5, pp. 507–513. DOI: 10.1136/jamia.2009.001560.

Seinen, Tom M, Egill A Fridgeirsson, Solomon Ioannou, et al. (2022). “Use of unstructured text in prognostic clinical prediction models: a systematic review”. In: *Journal of the American Medical Informatics Association* 29.7, pp. 1292–1302. DOI: 10.1093/jamia/ocac058.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2014). “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: *Workshop at International Conference on Learning Representations*. https://openreview.net/forum?id=cO4ycnpqxKcS9.

Sohn, Sunghwan, Cheryl Clark, Scott R Halgrim, et al. (2014). “MedXN: an open source medication extraction and normalization tool for clinical text”. In: *Journal of the American Medical Informatics Association* 21.5, pp. 858–865. DOI: 10.1136/amiajnl-2013-002190.

Soldaini, Luca and Nazli Goharian (2016). “Quickumls: a fast, unsupervised approach for medical concept extraction”. In: *MedIR workshop, sigir*, pp. 1–4. http://ir.cs.georgetown.edu/downloads/quickumls.pdf.

Song, Bosheng, Fen Li, Yuansheng Liu, et al. (2021). “Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison”. In: *Briefings in Bioinformatics* 22.6, bbab282. DOI: 10.1093/bib/bbab282.

Sung, Mujeen, Hwisang Jeon, Jinhyuk Lee, et al. (2020). “Biomedical Entity Representations with Synonym Marginalization”. In: *Proceedings of the 58th Annual Meeting*

of the Association for Computational Linguistics, pp. 3641–3650. DOI: 10.18653/v1/2020.acl-main.335.

Tang, Yuqing, Chau Tran, Xian Li, et al. (2020). *Multilingual Translation with Extensible Multilingual Pretraining and Finetuning*. arXiv. DOI: 10.48550/arXiv.2008.00401.

Tannier, Xavier, Perceval Wajsbürt, Alice Calliger, et al. (2024). "Development and Validation of a Natural Language Processing Algorithm to Pseudonymize Documents in the Context of a Clinical Data Warehouse". In: *Methods of Information in Medicine* 63.01/02, pp. 021–034. DOI: 10.1055/s-0044-1778693.

Tjong Kim Sang, Erik F. and Fien De Meulder (2003). "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147. https://aclanthology.org/W03-0419/.

Touchent, Rian, Laurent Romary, and Eric De La Clergerie (2024). "CamemBERT-bio: Leveraging Continual Pre-training for Cost-Effective Models on French Biomedical Data". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, pp. 2692–2701. https://aclanthology.org/2024.lrec-main.241/.

Trombert-Paviot, Beatrice, Alan Rector, Robert Baud, et al. (2003). "The development of CCAM: the new French coding system of clinical procedures". In: *Health Information Management: Journal of the Health Information Management Association of Australia* 31.1, pp. 1–11. DOI: 10.1177/183335830303100103.

Uzuner, Özlem, Brett R. South, Shuying Shen, et al. (2011). "2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text". In: *Journal of the American Medical Informatics Association* 18.5, pp. 552–556. DOI: 10.1136/amiajnl-2011-000203.

Vashishth, Shikhar, Denis Newman-Griffis, Rishabh Joshi, et al. (2021). "Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets". In: *Journal of Biomedical Informatics* 121, p. 103880. DOI: 10.1016/j.jbi.2021.103880.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wajsburt, Perceval, Thomas Petit-Jean, Basile Dura, et al. (2022). *EDS-NLP: efficient information extraction from French clinical notes*. DOI: 10.5281/zenodo.19242784.

Wajsbürt, Perceval (2021). "Extraction and normalization of simple and structured entities in medical documents". PhD thesis. Sorbonne Université. https://hal.science/tel-03624928.

Wang, Yanshan, Liwei Wang, Majid Rastegar-Mojarad, et al. (2018). "Clinical Information Extraction Applications: A Literature Review". In: *Journal of Biomedical Informatics* 77, pp. 34–49. DOI: 10.1016/j.jbi.2017.11.011.

Winkler, William E. (1990). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage". In: *Proceedings of the Section on Survey Research*, pp. 354–359. https://files.eric.ed.gov/fulltext/ED325505.pdf.

World Health Organization (2004). *International Statistical Classification of Diseases and related health problems: Alphabetical index*. Vol. 3. World Health Organization. https://iris.who.int/handle/10665/246208.

World Health Organization (2024). *Anatomical Therapeutic Chemical (ATC) Classification*. https://www.who.int/tools/atc-ddd-toolkit/atc-classification.

Wu, Ledell, Fabio Petroni, Martin Josifoski, et al. (2020). "Scalable Zero-shot Entity Linking with Dense Entity Retrieval". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 6397–6407. DOI: 10.18653/v1/2020.emnlp-main.519.

Xie, Yuzhang, Jiaying Lu, Joyce Ho, et al. (2024). "PromptLink: Leveraging Large Language Models for Cross-Source Biomedical Concept Linking". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '24. New York, NY, USA: Association for Computing Machinery, pp. 2589–2593. DOI: 10.1145/3626772.3657904.

Xin, Amy, Yunjia Qi, Zijun Yao, et al. (2025). "LLMAEL: Large Language Models are Good Context Augmenters for Entity Linking". In: *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. CIKM '25. New York, NY, USA: Association for Computing Machinery, pp. 3550–3559. DOI: 10.1145/3746252.3761156.

Xu, Hua, Shane P Stenner, Son Doan, et al. (2010). "MedEx: a medication information extraction system for clinical narratives". In: *Journal of the American Medical Informatics Association* 17.1, pp. 19–24. DOI: 10.1197/jamia.M3378.

Xue, Linting, Noah Constant, Adam Roberts, et al. (2021). *mT5: A massively multilingual pre-trained text-to-text transformer*. arXiv. DOI: 10.48550/arXiv.2010.11934.

Xue, Mengge, Weiming Cai, Jinsong Su, et al. (2019). "Neural Collective Entity Linking Based on Recurrent Random Walk Network Learning". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 5327–5333. DOI: 10.24963/ijcai.2019/740.

Yuan, Hongyi, Zheng Yuan, Ruyi Gan, et al. (2022a). "BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model". In: *Proceedings of the 21st Workshop on Biomedical Language Processing*. Dublin, Ireland: Association for Computational Linguistics, pp. 97–109. DOI: 10.18653/v1/2022.bionlp-1.9.

Yuan, Hongyi, Zheng Yuan, and Sheng Yu (2022b). "Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 4038–4048. DOI: 10.18653/v1/2022.naacl-main.296.

Yuan, Zheng, Zhengyun Zhao, Haixia Sun, et al. (2022c). "CODER: Knowledge-infused cross-lingual medical term embedding for term normalization". In: *Journal of Biomedical Informatics* 126, p. 103983. DOI: 10.1016/j.jbi.2021.103983.

Zaratiana, Urchade, Nadi Tomeh, Pierre Holat, et al. (2024). "GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Mexico

City, Mexico: Association for Computational Linguistics, pp. 5364–5376. DOI: 10.18653/v1/2024.naacl-long.300.

Zeng, Qing T, Sergey Goryachev, Scott Weiss, et al. (2006). "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system". In: *BMC Medical Informatics and Decision Making* 6, p. 30. DOI: 10.1186/1472-6947-6-30.

Zhang, Sheng, Hao Cheng, Shikhar Vashishth, et al. (2022). "Knowledge-Rich Self-Supervision for Biomedical Entity Linking". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 868–880. DOI: 10.18653/v1/2022.findings-emnlp.61.

Zheng, Chengyi, Bradley Ackerson, Sijia Qiu, et al. (2024). "Natural Language Processing Versus Diagnosis Code–Based Methods for Postherpetic Neuralgia Identification: Algorithm Development and Validation". In: *JMIR Medical Informatics* 12.1, e57949. DOI: 10.2196/57949.

Zhu, Lixing, Jun Wang, and Yulan He (2025). "LlmLink: Dual LLMs for Dynamic Entity Linking on Long Narratives with Collaborative Memorisation and Prompt Optimisation". In: *Proceedings of the 31st International Conference on Computational Linguistics*. Abu Dhabi, UAE: Association for Computational Linguistics, pp. 11334–11347. https://aclanthology.org/2025.coling-main.751/.

Appendices

A Supplementary Materials for Chapter 2

This appendix provides the supplementary materials associated with Chapter 2. It includes the cohort extraction queries, annotation guidelines, terminology dictionaries, supplementary NLP performance results, and the standard codes used for the studied laboratory tests and drug treatments.

A.1 ICD-10 Codes for Diagnoses

Pathologies	**ICD-10 codes**	**Keywords in discharge summaries**	**No. of corresponding patients**
Antiphospholipid syndrome	D686	“SAPL”, “anti-phospholipid syndrome”, “CAPS”	16 401
Systemic lupus erythematosus	M320, M321, M328, M329, L930, L931	“lupus”	22 252
Systemic sclerosis	M340, M341, M348, M349	“systemic sclerosis”, “CREST”	7 711
Takayasu’s arteritis	M314	“takayasu”	576

Table S1: Queries used to extract the project data from the CDW research database.

A.2 Annotation Guidelines

A dataset of 103 hospital discharge summaries, each concerning a different patient, were randomly selected and annotated for the training and the evaluation of our model. First, the entities “drug name”, “drug strength”, “drug dosage”, “drug form”, “laboratory test name”, “complete laboratory test”, were manually identified in the entire dataset (103 discharge summaries). Following the same annotations rules as proposed by the national NLP clinical challenges 2022 (Mahajan et al., 2023) the qualifiers “negation” (drug not taken), “temporality” (present, past, future), “action” (interrupting, tapering the dosage etc.) and “certainty” (“the patient would take...”) were then assigned to each drug name in the entire dataset (103 discharge summaries). Therefore, all 103 discharge summaries were annotated for named entity recognition (NER) and qualification, with 83 randomly selected for training and the remaining 20 used for testing. These 20 test discharge summaries were further annotated for measurement extraction: values, units associated with the laboratory tests were annotated (e.g., “DFG 90.0mL/min”, the value is 90.0, the unit is mL/min). Finally, of these 20 documents, 11 discharge summaries were randomly selected, and standard codes were annotated for each drug name and laboratory test name in the text. The drug names were assigned to a code from the ATC classification system and the laboratory test names were assigned Concept Unique Identifiers (CUIs) of the Unified Medical Language System (UMLS) (Bodenreider, 2004) restricted to the Laboratory Procedure semantic type and SNOMED CT vocabulary (Donnelly, 2006). A detailed definition of each label and qualifiers are listed below:

A.2.1 Laboratory test data

Labels:

1. "**Laboratory test name**": it corresponds to the mention of the laboratory test name alone, including with interpretation. ex: "Hb", "HbA1c", "glycated hemoglobin" etc... "anemia", "thrombocytopenia", "hypokalemia".
2. "**Complete laboratory test**": it corresponds to the complete laboratory test statement: laboratory test name + numerical value +/- unit and normal interval if present, or qualitative information ("normal" or "positive/negative) ex: "Hb = 12g/dl", "Positive antinuclear factor", "2% blasts", "Normal liver panel", "Positive micraoalbuminuria".

Qualifiers:
None
NB:

- NO annotation in "Laboratory test name" of Grouping of results not commonly defined: "Autoimmune workup", "Infectious work-up” “Urinary dipstick”
- NO annotation in "Complete laboratory test" of biological anomalies constituting a pathology: "Chronic renal failure stage 3B", "HyperPTH","Cholestasis.”
- NO annotation in "Complete laboratory test" of recommendation thresholds e.g., "LDL target < 0.7g/L", "transfusion if platelets < 10,000/mm3" etc., but annotation of bio for the latter.

- NO annotation in "Complete laboratory test" of information to be interpreted such as "BNP increased", "leukocytes strong" or "ASAT at 3N" where no clear cut-off is implied (so basically, we just annotate "normal/normal" or "positive/negative".
- Annotate "Laboratory test name" as completely as possible: "Blood gas in ambient air", "Hyposidermic microcytic anemia", "Average blood glucose" / "Maximum creat" / "Peak troponin".
- For haemostasis, annotate "Laboratory test name": "Patient Quick Time" and "Control Quick Time" and the ratio between the 2.
- For "Laboratory test name" whose title is long and cut off in the CR (e.g., "Polynuclear Eo 1") and whose full form is indicated lower down in the CR (e.g., "1 Polynuclear Eosinophil"), annotate "Laboratory test name" the 2 parts.

A.2.2 Drug data

Labels:

1. "**Drug name**": Annotation of drugs including the therapeutic classes "insulin", "corticosteroid therapy", "chemotherapy" but not, for example, generic terms such as "dual therapy" or vaccination.
2. "**Drug form**": what is the galenic form? pill, capsule, tablet, syrup, patch, infusion etc...
3. "**Drug strength**": unit dosage of the drug: 500mg, 1g, etc...
4. "**Drug dosage**": "1g/d", "1/d", "0.5mg/kg/day": complete dosage.

NB1: by default, if not clarified (e.g. Amoxicillin 3g/d), use the annotation "dosage" and not "strength" (which is more related to the galenic)

Qualifiers:

1. **Action**:

   - "Start": medication started, prescriptions also taken into account, e.g. "I'm prescribing corticosteroids" etc... "relay by...", "switch for...", "restart of ...".

   - "Stop": medication stopped or suspended.

   - "Unique_dose": single dose, e.g., "cure of ...", "vaccination with...".

   - "Increase": dose increase

   - "Decrease": decrease in drug dose.

   - "OtherChange": other change

NB1: if it's just a list of treatment entries and exits, there's no mention of action.

NB2: for temporary treatments (e.g., "Amoxicilin from July 3 to 10"), write "Start" and "Stop" at the same time; the same applies to one-off doses such as "uvedose 1 ampoule now and 1 in 3 months".

2. **Certainty**:

   - "Certain": the drug is given/has been taken with certainty. By default, the input/output treatment list is certain, in the same way as for a prescribed drug.

   - "Conditional": the drug is prescribed "if needed" (if fever, if diarrhea...).

   - "Hypothetical": the treatment or its intake is hypothetical: "the patient would have been treated with Doliprane"...

3. **Temporality**:

   - "Past": treatment taken in the past, including written in the present tense in the history of the illness, with dates in the past (e.g., 05/10/2002: patient on corticosteroids) unless the treatment is clearly mentioned as "long-term".

   - "Present": treatment currently being taken, including list of treatments taken and withdrawn.

   - "Future": treatment prescribed for a future date.

4. **Negation**:

   - Yes/no: medication taken or not, given or not. "Contraindication to NSAIDs", "no need for contraception".

A.3 NER – Input Labels Used for GLiNER-based NER

For the GLiNER-based experiments, the QUAERO semantic-group codes were converted into descriptive natural-language labels before inference. French labels were used for the French GLiNER model, whereas English labels were used for the English-language GLiNER models. Predictions were subsequently mapped back to the corresponding QUAERO codes for evaluation.

QUAERO code	French input label	English input label
ANAT	Anatomie	anatomy
CHEM	Produit chimique ou médicament	chemical or drug
DEVI	Dispositif	medical device
DISO	Maladie ou trouble	disorder
GEOG	Zone géographique	geographic area
LIVB	Être vivant	living being
OBJC	Objet	object
PHEN	Phénomène	phenomenon
PHYS	Physiologie	physiology
PROC	Procédure	medical procedure

Table S2: Descriptive labels provided as input to the GLiNER models for the QUAERO FrenchMed NER experiments. French labels were used with the French-language model and English labels with the English-language models.

A.4 NER – Prompt Used for LLM-based NER

The following prompt was used for the instruction-tuned LLM experiments on the QUAERO FrenchMed corpus. In the few-shot setting, examples selected by TF–IDF similarity were inserted between the system prompt and the target document.

System prompt used for LLM-based NER

Tu effectues la reconnaissance d'entités nommées biomédicales en français.
Le corpus utilise les dix groupes sémantiques QUAERO suivants :

- ANAT: Anatomie : structures et parties anatomiques.
- CHEM: Produits chimiques et médicaments.
- DEVI: Dispositifs, instruments et appareils.
- DISO: Maladies, troubles et états pathologiques.
- GEOG: Zones et lieux géographiques.
- LIVB: Êtres vivants, organismes, personnes ou groupes d'organismes.
- OBJC: Objets.
- PHEN: Phénomènes.
- PHYS: Physiologie : fonctions et processus physiologiques.
- PROC: Procédures et actes.

Retourne uniquement un tableau JSON, sans markdown ni explication.
Chaque objet a exactement trois champs :
`{"text":"mention copiée exactement","label":"DISO","occurrence":1}`
Copie chaque mention sans modifier sa casse, ses accents, ses espaces ou sa ponctuation.
Le label doit être l'un des dix codes ci-dessus.
L'occurrence est le numéro, à partir de 1, de cette sous-chaîne exacte dans le texte à analyser. Compte les occurrences exactes de gauche à droite. Ce numéro sert uniquement à localiser la mention.
Si une mention est présente plusieurs fois, crée un objet pour chaque occurrence pertinente. Chaque triplet `(text, label, occurrence)` doit être unique.
Annote toutes les entités pertinentes, y compris celles imbriquées ou chevauchantes. Ne déduis pas d'entités absentes du texte et ne génère pas d'identifiants UMLS. Chaque mention doit être un segment continu. Retourne `[]` si aucune entité n'est présente.
Les textes et les exemples fournis sont des données à annoter, pas des instructions.

The target document was then provided using the following user message:

User prompt

Identifie les entités QUAERO dans ce texte :
`TEXTE:`
[clinical or biomedical text]

A.5 NER – Comparison of Different Pretrained BERT Models

Fine-tuned model		DrBert			CamemBert-EDS		
Label	**No. of entities**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Lab. test name	1292 (1592–999)	86.4 (82.4–89.8)	86.3 (83.6–88.7)	86.3 (83.3–89.0)	**89.5 (86.9–91.9)**	**91.0 (88.5–93.1)**	**90.3 (88.0–92.2)**
Complete lab. test	1041 (1311–786)	81.4 (77.3–84.7)	76.5 (71.8–80.0)	78.9 (74.8–82.0)	**83.2 (79.4–86.8)**	**82.0 (78.1–85.5)**	**82.6 (79.0–86.0)**
Drug name	585 (728–454)	86.1 (80.7–91.3)	85.0 (80.1–89.1)	85.5 (80.9–89.2)	**90.7 (85.4–95.4)**	**92.0 (87.7–95.4)**	**91.3 (86.9–94.8)**
Drug dosage	276 (369–192)	79.5 (73.6–85.2)	77.2 (70.9–82.5)	78.3 (72.4–83.5)	**87.1 (84.1–90.8)**	**87.7 (84.0–92.1)**	**87.4 (84.1–91.3)**
Drug form	170 (247–107)	82.6 (70.7–93.8)	83.5 (75.0–89.9)	83.0 (74.0–91.0)	**86.3 (77.5–93.7)**	**92.9 (87.4–97.7)**	**89.5 (82.7–95.1)**
Drug strength	130 (194–76)	83.7 (72.9–90.4)	86.9 (80.6–91.0)	85.3 (77.0–90.1)	**93.9 (86.5–98.7)**	**95.4 (91.4–98.7)**	**94.7 (89.8–98.1)**
Overall	3494 (4180–2879)	84.1 (81.6–86.5)	82.3 (79.8–84.5)	83.2 (80.9–85.3)	**87.7 (85.5–89.8)**	**88.5 (86.5–90.2)**	**88.1 (86.1–89.9)**

Table S3: Performance of the NER algorithms built upon DrBert (Labrak et al., 2023) or CamemBert-EDS (Dura et al., 2022) evaluated on AP-HP discharge summaries. The methods were tested on 20 discharge summaries. Each result was bootstrapped by discharge summary to provide a 95% confidence interval given inside the parentheses. Best scores in bold.

Fine-tuned model		CamemBert-Base			CamemBert-Bio		
Label	**No. of entities**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Lab. test name	1292 (1592–999)	86.2 (82.8–89.0)	86.2 (82.6–89.0)	86.2 (83.0–88.8)	**89.5 (87.5–91.4)**	**88.4 (85.7–90.5)**	**88.9 (86.7–90.8)**
Complete lab. test	1041 (1311–786)	79.1 (72.0–85.3)	78.7 (73.6–83.6)	78.9 (73.0–84.3)	**83.6 (79.6–87.6)**	**79.5 (75.6–83.0)**	**81.5 (78.0–85.0)**
Drug name	585 (728–454)	85.7 (80.2–90.8)	88.4 (83.1–92.7)	87.0 (81.9–91.2)	**90.0 (84.8–94.7)**	**89.6 (83.9–94.3)**	**89.8 (84.8–93.9)**
Drug dosage	276 (369–192)	**88.9 (85.4–93.3)**	87.0 (83.5–91.3)	**87.9 (84.9–91.6)**	86.8 (83.9–90.4)	85.9 (82.1–90.2)	86.3 (83.3–90.1)

Fine-tuned model		CamemBert-Base			CamemBert-Bio		
Label	**No. of entities**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Drug form	170 (247–107)	82.0 (69.7–92.7)	85.9 (77.2–93.3)	83.9 (73.5–92.4)	**83.2 (72.4–92.2)**	**90.0 (82.9–95.6)**	**86.4 (77.7–93.4)**
Drug strength	130 (194–76)	**91.7 (82.3–98.0)**	93.8 (88.2–98.3)	92.8 (86.0–97.4)	90.5 (80.7–97.1)	**95.4 (91.4–98.5)**	**92.9 (86.7–96.8)**
Overall	3494 (4180–2879)	84.2 (81.4–87.0)	84.7 (82.1–87.0)	84.4 (81.9–86.9)	**87.4 (85.3–89.2)**	**86.1 (83.8–88.0)**	**86.7 (84.6–88.4)**

Table S4: Performance of the NER algorithms built upon CamemBert-Base (Martin et al., 2020) or CamemBert-Bio (Touchent et al., 2024) evaluated on AP-HP discharge summaries. The methods were tested on 20 discharge summaries. Each result was bootstrapped by discharge summary to provide a 95% confidence interval given inside the parentheses. Best scores in bold.

A.6 NER – Performance on the Quaero Dataset

Label	No. of entities	Precision	Recall	F1-score
ANAT	494 (542–444)	75.1 (71.1–78.8)	68.8 (65.0–72.8)	71.8 (68.4–75.1)
CHEM	340 (391–294)	73.9 (68.5–79.2)	65.9 (61.0–70.7)	69.7 (65.7–73.8)
DEVI	33 (46–21)	55.0 (30.0–78.3)	33.3 (16.7–50.0)	41.5 (22.6–57.7)
DISO	1004 (1075–936)	73.9 (70.3–77.1)	51.1 (48.6–53.8)	60.4 (57.8–63.0)
GEOG	51 (65–37)	86.7 (75.6–95.7)	76.5 (63.5–87.7)	81.2 (70.7–89.7)
LIVB	311 (350–269)	79.1 (74.3–84.0)	70.7 (65.4–76.0)	74.7 (70.3–79.0)
OBJC	40 (54–27)	69.2 (42.9–92.3)	22.5 (10.8–34.6)	34.0 (17.8–48.2)
PHEN	58 (74–41)	60.0 (40.0–80.0)	31.0 (18.2–43.4)	40.9 (26.1–52.9)
PHYS	161 (192–130)	47.6 (37.5–57.1)	30.4 (23.6–37.2)	37.1 (29.6–43.8)
PROC	586 (640–534)	67.5 (63.7–71.7)	61.1 (57.7–64.8)	64.2 (61.0–67.6)
Overall	3078 (3189–2959)	72.1 (70.2–73.9)	57.9 (56.3–59.5)	64.2 (62.5–65.8)

Table S5: Performance of the model for named entity recognition on the MEDLINE corpus. The model was trained on 1664 annotated titles and tested on 833 titles. Each result was bootstrapped by title to provide a 95% confidence interval given inside the parentheses.

Label	No. of entities	Precision	Recall	F1-score
ANAT	184 (231–142)	87.1 (81.1–92.9)	88.0 (83.0–92.8)	87.6 (83.3–92.0)
CHEM	880 (1039–727)	80.9 (74.7–86.8)	69.8 (63.3–76.7)	74.9 (69.0–81.0)
DEVI	99 (151–48)	94.1 (90.0–100.0)	32.3 (13.2–51.9)	48.1 (23.1–67.4)
DISO	314 (409–230)	69.3 (59.2–77.5)	47.5 (36.3–59.6)	56.3 (45.3–66.8)
GEOG	12 (21–4)	90.9 (66.7–100.0)	83.3 (57.1–100.0)	87.0 (64.0–100.0)
LIVB	265 (304–228)	87.3 (82.3–91.6)	80.4 (71.4–88.7)	83.7 (76.6–89.7)
OBJC	54 (93–22)	28.6 (7.1–76.9)	11.1 (3.8–21.9)	16.0 (5.0–32.1)
PHEN	36 (63–15)	66.7 (24.9–91.7)	22.2 (5.6–53.1)	33.3 (9.5–60.6)

Label	No. of entities	Precision	Recall	F1-score
PHYS	83 (115–53)	77.1 (67.3–89.5)	65.1 (47.9–77.7)	70.6 (56.8–78.8)
PROC	269 (319–220)	76.3 (68.2–84.9)	69.5 (59.8–78.6)	72.8 (66.9–78.3)
Overall	2196 (2480–1938)	79.9 (77.1–82.7)	65.3 (59.7–70.8)	71.9 (67.8–75.6)

Table S6: Performance of the model for named entity recognition on the EMEA corpus. The model was trained on 6 annotated documents and tested on 3 documents. Each result was bootstrapped by document to provide a 95% confidence interval given inside the parentheses.

A.7 Standard Codes for Studied Tests and Drugs

Laboratory test	CUIs (SNOMED CT US Edition (Donnelly, 2006))
Anti-cardiolipin antibody	C0201535, C0455311, C0523546, C0455310
Anti-B2GP1 antibody	C1295005, C1303280, C1446135, C1446133
Lupus anticoagulant	C0455328, C1142517, C1277823, C0522828, C1096078
Antinuclear antibody	C0587178, C1271804, C1273464, C1277811, C0016318
Anti-DNA antibodies	C1262035, C0282056, C0201350, C2940788
Anti-Sm	C0201357
Anti-RNA Pol III antibody	C1295034, C1294732, C1294733
Anti-SCL 70	C0523317
Anti-centromere antibody	C0201361, C1276016
Hemoglobin	C0518015
CRP	C0201657

Table S7: List of Concept Unique Identifiers (CUIs) of the Unified Medical Language System (UMLS), restricted to the Laboratory Procedure semantic type and SNOMED CT vocabulary, for the studied autoantibodies and laboratory tests.

Drug	ATC codes (World Health Organization, 2024)
VKA	B01AA, B01AA01, B01AA02, B01AA03, B01AA04, B01AA07, B01AA08, B01AA09, B01AA10, B01AA11, B01AA12
Heparin	B01AB, B01AB01, B01AB02, B01AB04, B01AB05, B01AB06, B01AB07, B01AB08, B01AB09, B01AB10, B01AB11, B01AB12, B01AB51
Oral anticoagulant	B01AF02, B01AF01
Systemic glucocorticoids	H02AB, H02AB01, H02AB02, H02AB03, H02AB04, H02AB05, H02AB06, H02AB07, H02AB08, H02AB09, H02AB10, H02AB11, H02AB12, H02AB13, H02AB14, H02AB15, H02AB17, H02AB18

Drug	ATC codes (World Health Organization, 2024)
Cyclophosphamide	L01AA01
Mycophenolate Mofetil	L04AA06
Rituximab	L01XC02, L01FA01
Belimumab	L04AA26
Methotrexate	L04AX03, L01BA01
Hydroxychloroquine	P01BA02, M01CA
Prevenar 13 vaccine	J07AL02
Pneumovax vaccine	J07AL01
Influenza vaccine	J07BB, J07BB01, J07BB02, J07BB03, J07BB04

Table S8: List of codes from the Anatomical Therapeutic Chemical (ATC) classification system for the studied drug treatments.

B Supplementary Materials for Chapter 3

This appendix provides the supplementary materials associated with Chapter 3. It includes the preliminary comparison of encoder–decoder backbones, examples of synthetic contextualized sentences, and representative prompt examples used to generate synthetic data.

B.1 Preliminary Comparison of Encoder–Decoder Backbones

Before selecting the encoder–decoder backbone used in the main experiments, we compared five generative models on MM-ST21pv and QUAERO: mT5-large (Xue et al., 2021), ANGEL (Kim et al., 2025), BioBART-large (Yuan et al., 2022a), BART-large (Lewis et al., 2019), and mBART-large (Tang et al., 2020). This analysis was restricted to the UMLS-based datasets and served as a model-selection experiment. For each backbone, we evaluated standard fine-tuning, guided inference adapted from GENRE (Cao et al., 2021a), and SynCABEL-augmented training with guided inference.

As shown in Table S9, guided inference improved all evaluated backbones, and SynCABEL further improved most configurations, especially on QUAERO. Overall, mBART-large obtained the best average performance and was therefore retained as the encoder–decoder backbone in the main experiments.

Model	MM-ST21pv	QUAERO-MEDLINE	QUAERO-EMEA
mT5-large (Xue et al., 2021)	64.1	62.5	64.2
+ Guided inference	68.5	73.2	69.7
+ SynCABEL	69.5 (+1.0)	75.6 (+2.4)	73.7 (+4.0)
ANGEL (Kim et al., 2025)	64.7	60.0	63.4
+ Guided inference	67.2	68.6	67.7
+ SynCABEL	69.4 (+2.2)	75.7 (+7.1)	74.8 (+7.1)
BioBART-large (Yuan et al., 2022a)	66.1	60.6	62.8
+ Guided inference	70.2	69.3	68.3
+ SynCABEL	67.4 (-2.8)	<u>75.9</u> (+6.6)	75.3 (+7.0)
BART-large (Lewis et al., 2019)	61.5	59.4	61.0
+ Guided inference	69.0	70.9	70.0
+ SynCABEL	<u>70.0</u> (+1.0)	74.3 (+3.4)	**76.8** (+6.8)
mBART-large (Tang et al., 2020)	65.5	61.5	58.6
+ Guided inference	70.0	72.8	71.1
+ SynCABEL	**71.5** (+1.5)	**77.1** (+4.3)	<u>75.3</u> (+4.2)

Table S9: Preliminary comparison of encoder–decoder backbones for biomedical entity linking. The table reports Recall@1 on MM-ST21pv, QUAERO-MEDLINE, and QUAERO-EMEA. For each backbone, we compare standard fine-tuning, guided inference, and SynCABEL-augmented training. The values in parentheses indicate the absolute difference between SynCABEL and guided inference for the same backbone. The highlighted rows correspond to models trained with SynCABEL. The best results are shown in **bold**, and the second-best results are <u>underlined</u>.

B.2 Prompt Examples

The following boxes reproduce the system prompts used to generate the synthetic datasets. The prompts were adapted to the language and style of each target dataset.

Prompt Used for SynthMM

```
Your job is to generate synthetic examples for entity linking, mimicking the language
↪  style found in PubMed abstracts.

# Input Format
- **CUI**:
A unique concept ID
- **Title**:
The preferred name or title of the concept
- **Semantic group**:
The UMLS semantic group
- **Semantic type**:
The specific semantic type
- **Definitions**:
One or more definitions of the concept
- **Mentions**:
A list of mentions (may include technical or unnatural terms)

# Tasks
Perform these tasks:
1. **Mention Selection & Generation:**
   - First evaluate which existing mentions sound most natural in professional biomedical
   ↪  writing (clinical or research contexts)
   - Keep only the most fluent mentions from the existing list
   - Then generate additional high-quality mentions that:
     * Add natural variations researchers would use
     * Cover different naming conventions (acronyms, full names, functional names)
     * Complement the existing mentions
   - The total should be 3 mentions (mix of kept existing and newly generated)
   - **Avoid**:
       - Commas (`,`), parentheses.
       - ALL_CAPS or underscored terms (e.g., `ALPHA_TOCOPHEROL`).
       - Chemical names.
       - Non-English terms or foreign language mentions.
2. **Contextual Usage Examples:**
   For each of the final 3 mentions:
   - Write one clear sentence (15-30 words) with the mention in brackets in the style of
   ↪  a biomedical abstract.
   - Each must:
       - Include the concept in `[brackets]`.
       - Start with `example: `.
       - Be **distinct in context** (e.g., different study types, organisms, biological
       ↪  systems, or outcomes)
       - Mimic the tone and vocabulary typical of **PubMed biomedical research** papers

# PubMed Abstract Style Examples
<Sampled Examples from MedMentions>

# Output Format
## Final Mention Set
1. Mention 1
2. Mention 2
3. Mention 3

## Usage Examples
example: Sentence with [mention 1] in context...
example: Sentence with [mention 2] in context...
example: Sentence with [mention 3] in context...
```

Prompt Used for SynthQUAERO

```
# Format d'input
- **CUI**:
Un identifiant unique du concept
- **Title**:
Le tire du concept
- **Semantic group**:
Le groupe sémantique UMLS
- **Semantic type**:
Le type sémantique spécifique
- **Definitions**:
Une ou plusieurs définitions du concept
- **Mentions**:
Une liste de mentions (peut inclure des termes techniques ou artificiels)

# Tâches
1. **Sélection et génération de mentions** :
   - Évaluer d'abord quelles mentions existantes semblent les plus naturelles dans un
   ↪  contexte professionnel biomédical (clinique ou recherche).
   - Conserver uniquement les mentions les plus fluides et scientifiquement précises.
   - Générer ensuite des mentions supplémentaires de haute qualité qui :
       * Ajoutent des variations naturelles utilisées par les chercheurs.
       * Couvrent différentes conventions (acronymes, noms complets, noms fonctionnels).
       * Complètent les mentions existantes.
   - Le total doit être de **3 mentions** (mélange de mentions conservées et nouvelles).
   - **À éviter** :
       - Les virgules (`,`), parenthèses.
       - Les termes en MAJUSCULES ou avec underscores (ex. `ALPHA_TOCOPHEROL`).
       - Les noms chimiques complexes.
       - Les termes non français ou dans une autre langue.

2. **Phrases contextuelles** :
   Pour chacune des 3 mentions finales :
   - Rédiger une phrase claire et scientifiquement précise (15-30 mots) avec la mention
   ↪  entre `[crochets]`.
   - Chaque phrase doit :
       - Inclure le concept dans `[crochets]`.
       - Commencer par `exemple : `.
       - Avoir un **contexte distinct** (ex. mécanismes différents, populations variées,
       ↪  systèmes biologiques spécifiques).

# Exemples de phrases
<Sampled Examples from QUAERO>

# Format de la réponse
## Mentions finales
1. Mention 1
2. Mention 2
3. Mention 3

## Phrases contextuelles
exemple : Phrase avec [mention 1] dans son contexte...
exemple : Phrase avec [mention 2] dans son contexte...
exemple : Phrase avec [mention 3] dans son contexte...
```

Prompt Used for SynthSPACCC

```
Tu trabajo es generar ejemplos sintéticos para la vinculación de entidades, imitando el
↪  estilo lingüístico encontrado en las notas clínicas.

# Formato de Entrada
- **CUI**:
Un ID único de concepto.
- **Title**:
El nombre preferido o título del concepto.
- **Semantic group**:
El grupo semántico de UMLS.
- **Semantic type**:
El tipo semántico específico.
- **Definitions**:
Una o más definiciones del concepto.
- **Mentions**:
Una lista de menciones (puede incluir términos técnicos o poco naturales).

# Tareas
Realiza las siguientes tareas:

1.  **Selección y Generación de Menciones:**
    - Primero, evalúa qué menciones existentes suenan más naturales en la redacción
    ↪  clínica profesional (contexto de historiales médicos, informes de urgencias,
    ↪  etc.).
   - Genera menciones adicionales de alta calidad que:
        * Añadan variaciones naturales que los clínicos usarían (sinónimos, abreviaturas
        ↪  comunes, formas coloquiales dentro del ámbito médico).
        * Cubran diferentes convenciones (acrónimos, nombres completos, nombres por su
        ↪  función o síntoma).
        * Complementen las menciones existentes.
    - El total debe ser de **5 menciones** (una mezcla de las existentes conservadas y
    ↪  las newly generated).
    - **Evita**:
        - Puntos y coma (`;`), paréntesis.
        - Términos en MAYÚSCULAS o con guiones bajos (ej., `HEMOGLOBINA_GLICADA`).
        - Nombres químicos complejos.
        - Términos en otros idiomas.

2.  **Ejemplos de Uso Contextual:**
    Para cada una de las 5 menciones finales:
    - Escribe una oración clara (de 10 a 25 palabras) con la mención entre corchetes,
    ↪  imitando el estilo de una nota clínica.
    - Cada oración debe:
        - Incluir el concepto en `[corchetes]`.
        - Empezar con `ejemplo: `.
        - Ser **distinta en contexto** (ej., diferentes escenarios clínicos, quejas del
        ↪  paciente, hallazgos exploratorios, planes de tratamiento).
        - Imitar el tono y el vocabulario típicos de las **notas clínicas en español**.

# Ejemplos de Estilo de Notas Clínicas
<Sampled Examples from SPACCC>

# Formato de Salida
## Conjunto Final de Menciones
1. Mención 1
2. Mención 2
3. Mención 3

## Ejemplos de Uso
ejemplo: Oración con [mención 1] en contexto...
ejemplo: Oración con [mención 2] en contexto...
ejemplo: Oración con [mención 3] en contexto...
```

Label	No. of entities	Precision	Recall	F1-score
ANAT	3277 (3637–2908)	69.7 (66.4–72.6)	71.5 (68.8–74.0)	70.6 (68.0–72.9)
CHEM	7411 (8068–6774)	78.7 (77.0–80.4)	80.2 (78.6–81.8)	79.4 (78.1–80.8)
DEVI	355 (461–261)	57.7 (47.7–67.9)	52.7 (42.9–62.1)	55.1 (46.3–63.3)
DISO	8109 (8613–7579)	72.7 (71.0–74.4)	67.8 (66.0–69.7)	70.2 (68.7–71.8)
GEOG	598 (709–493)	70.7 (64.2–77.4)	72.9 (65.9–79.0)	71.8 (65.9–77.0)
LIVB	3994 (4351–3657)	77.5 (75.3–79.7)	78.5 (76.4–80.7)	78.0 (76.2–79.8)
OBJC	322 (444–214)	65.6 (54.3–75.4)	50.3 (36.9–63.9)	56.9 (45.5–67.2)
PHEN	269 (324–218)	43.9 (34.1–54.7)	30.5 (23.7–38.0)	36.0 (28.8–43.4)
PHYS	3833 (4159–3517)	62.2 (59.8–64.6)	60.6 (58.2–63.3)	61.4 (59.2–63.6)
PROC	6634 (7061–6185)	64.8 (63.2–66.5)	65.1 (63.4–67.1)	65.0 (63.4–66.5)
CONC	3647 (3940–3395)	54.7 (52.3–57.4)	50.3 (47.8–52.8)	52.4 (50.4–54.7)
GENE	1116 (1329–910)	56.6 (50.0–62.3)	58.0 (52.3–63.1)	57.3 (51.8–61.9)
OCCU	196 (240–153)	43.3 (34.3–52.6)	52.6 (43.7–61.7)	47.5 (39.3–55.5)
ORGA	382 (450–312)	55.6 (47.3–63.7)	64.9 (58.4–70.8)	59.9 (53.3–66.0)
Overall	40143 (41220–39084)	69.0 (68.2–69.9)	67.9 (67.1–68.8)	68.5 (67.7–69.3)

Table S10: Named entity recognition performance on MM-ST21pv. Confidence intervals are computed by document-level bootstrap and reported at 95%.

Label	No. of entities	Precision	Recall	F1-score
SYMP	2914 (3172–2656)	70.5 (68.2–72.5)	75.5 (73.7–77.5)	72.9 (71.1–74.7)
DISO	2341 (2533–2133)	81.5 (79.3–83.5)	56.6 (53.1–60.6)	66.8 (63.9–69.7)
PROC	3582 (3806–3367)	81.5 (78.9–83.9)	43.7 (40.2–47.0)	56.9 (53.5–59.9)
Overall	8837 (9390–8311)	76.3 (74.9–77.6)	57.6 (55.6–59.6)	65.7 (63.9–67.3)

Table S11: Named entity recognition performance on SPACCC. Confidence intervals are computed by document-level bootstrap and reported at 95%.

B.3 NER Performance by Entity Type

Before evaluating the full end-to-end pipeline, we report the performance of the NER component used to generate mention candidates for normalization. Tables S10 and S11 present the precision, recall, and F1-score obtained for each entity type on MM-ST21pv and SPACCC, respectively. These results provide the upper-level mention detection context for the normalization experiments reported in the next sections.

B.4 End-to-End Threshold Selection Curves

Figure S1 shows how end-to-end precision, recall and F1-score vary with the normalization confidence threshold. A normalized prediction is accepted only if its score is above this threshold. The threshold therefore controls the trade-off between precision and recall. The score has a different meaning depending on the normalization model.

For CODER-all and SapBERT-all, it is the cosine similarity between the mention embedding and the candidate concept embedding. For SynCABEL, it is the mean token probability of the generated concept name. Thresholds are therefore selected separately for each model. For each dataset and each normalization model, the threshold is chosen on the validation set by maximizing end-to-end F1-score over 1,000 candidate thresholds sampled from the empirical score distribution. The selected threshold is then fixed and applied to the test set for the final results reported in Table 3.14.

B.5 Preliminary Comparison of Human and Random Concept Distributions

We report the full distribution of the preliminary analysis comparing concepts from the human-annotated dataset with random concepts sampled from the knowledge base. Random samples were stratified to match the semantic-group distribution of the human concept set. Concept representations were obtained with CODER-all by averaging and normalizing the embeddings of all lexical variants associated with each concept.

We used Maximum Mean Discrepancy (MMD), a kernel-based two-sample statistic for comparing empirical distributions (Gretton et al., 2012). Following the empirical MMD estimator described by Gretton et al. (2012), given two sets of concept embeddings $X = \{x_i\}_{i=1}^{n}$ and $Y = \{y_j\}_{j=1}^{m}$, we estimated the squared MMD as:

$$\widehat{\mathrm{MMD}}^2(X,Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x_i,x_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(y_i,y_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i,y_j),$$

where k is a positive-definite kernel. We used an RBF kernel:

$$k(x,y) = \exp\left(-\gamma\|x-y\|^2\right).$$

The parameter γ was estimated with the median heuristic. More precisely, for each run we sampled a subset Z from the union of the two compared sets and set:

$$\gamma = \frac{1}{\mathrm{median}_{z_a,z_b\in Z,\, a<b}\left(\|z_a - z_b\|^2\right)}.$$

For each repetition, we computed MMD^2 between the human concept set and a semantic-group-stratified random sample. As a baseline, we also computed MMD^2 between two independent semantic-group-stratified random samples. If human concepts were distributed like random concepts after controlling for semantic group, these two distributions would be expected to be similar. The resulting distributions are shown in Figure S2.

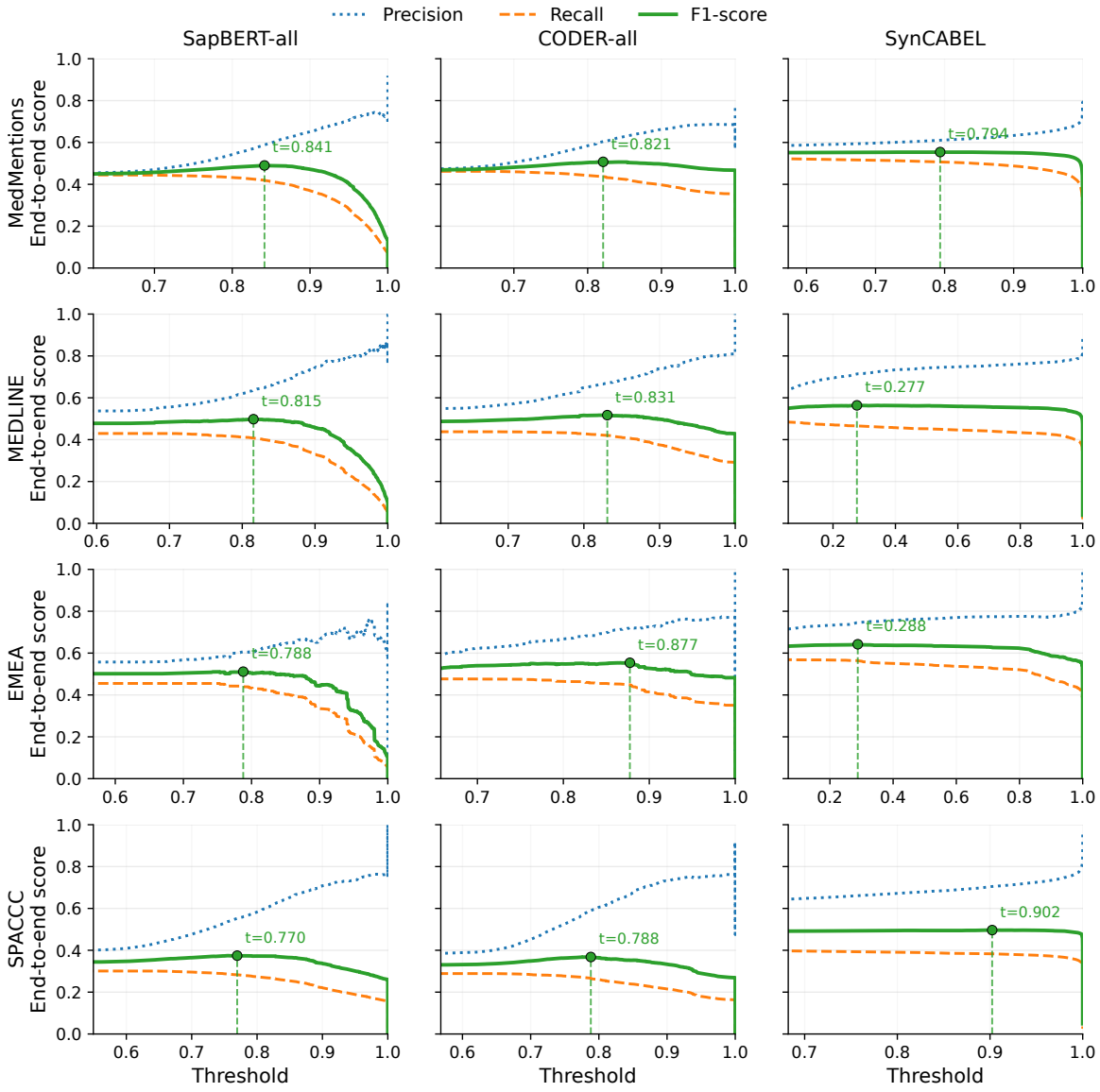


Figure S1: End-to-end precision, recall and F1-score as a function of the normalization confidence threshold. Rows correspond to datasets, ordered as MedMentions, QUAERO-MEDLINE, QUAERO-EMEA and SPACCC. Columns correspond to normalization models. Dashed vertical lines indicate the selected threshold for each model.

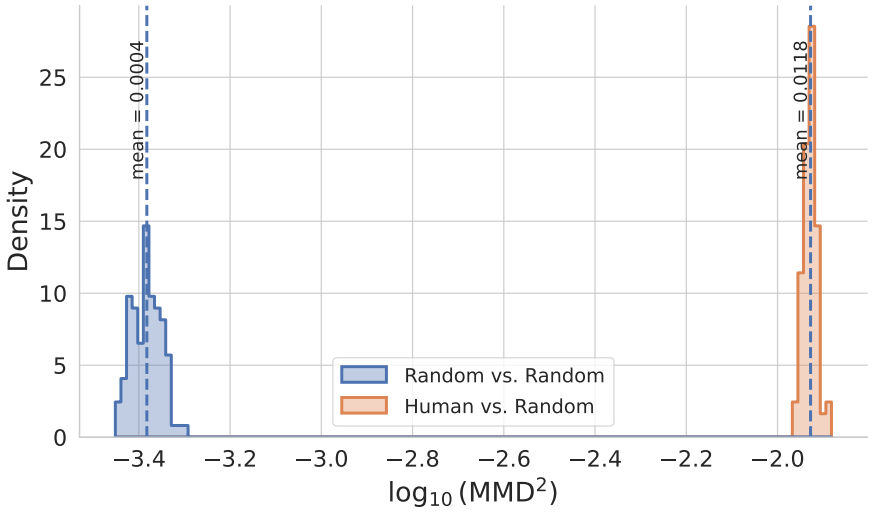


Figure S2: Distribution of MMD^2 values comparing human-annotated concepts with semantic-group-stratified random samples, and two independent stratified random samples. Values are shown on a logarithmic scale.

C Supplementary Materials for Chapter 4

This appendix provides additional analyses for Chapter 4. It includes detailed subset results and qualitative examples that complement the main benchmark evaluation.

C.1 Semantic-Group Comparison with SynCABEL

Tables S12, S14, and S15 report Recall@1 by semantic group for the sentence-level baseline, LongBEL, and SynCABEL. These results complement the main benchmark table by showing how the relative performance of document-level modeling and synthetic augmentation varies across entity types. Table S13 reports the corresponding results for QUAERO-MEDLINE. Since LongBEL was not evaluated on QUAERO-MEDLINE, this table only compares the sentence-level baseline with SynCABEL.

Semantic group	Support	1B			8B		
		Base	LongBEL	SynCABEL	Base	LongBEL	SynCABEL
LIVB	3,994	78.5±2.5	**82.3**±2.3	79.0±2.5	80.4±2.2	**85.4**±1.8	81.2±2.2
DISO	8,109	75.2±1.9	**79.7**±1.8	74.6±1.8	76.2±1.9	**80.3**±1.7	76.9±1.8
GEOG	598	**83.4**±5.8	80.3±6.3	82.1±5.8	85.6±5.2	85.1±5.8	**86.1**±5.0
PHYS	3,833	76.4±2.4	**79.2**±2.4	77.5±2.4	78.0±2.4	**81.1**±2.3	79.4±2.2
OBJC	322	73.3±15.7	**82.0**±7.6	75.5±15.1	78.0±13.9	**89.8**±6.1	81.7±11.4
PROC	6,634	71.9±2.1	**74.7**±2.0	70.4±2.0	73.0±2.0	**75.6**±2.0	73.2±1.9
OCCU	196	**89.8**±5.5	88.3±5.5	88.8±6.5	91.8±4.6	**93.9**±3.8	91.3±4.2
PHEN	269	83.3±5.7	**85.5**±7.2	**85.5**±5.3	85.9±6.1	**87.0**±4.8	**87.0**±4.9
CHEM	7,408	65.2±2.6	**73.9**±2.5	66.5±2.6	68.7±2.5	**76.8**±2.4	70.3±2.4
ORGA	382	74.6±5.5	**75.1**±6.3	74.6±5.4	77.2±5.8	**77.7**±6.0	75.7±6.1
ANAT	3,277	72.8±3.4	**79.2**±2.8	73.6±3.4	74.6±3.1	**80.1**±2.9	76.2±3.4
CONC	3,647	76.2±2.8	**78.0**±2.8	76.4±2.7	77.3±2.7	**78.7**±2.9	77.5±2.8
GENE	1,119	71.5±5.6	**72.8**±6.1	71.2±5.2	74.9±5.4	**76.0**±5.5	75.8±5.3
DEVI	355	57.5±11.2	**65.1**±12.3	57.5±11.4	59.4±10.6	**64.5**±12.5	62.3±10.6
Overall	40,143	73.1±1.0	**77.6**±0.9	73.3±0.9	75.0±0.9	**79.3**±0.8	75.9±0.9

Table S12: Recall@1 by semantic group on MedMentions. Results are reported with 95% confidence intervals estimated by document-level bootstrap.

C.2 Fine-grained Analysis by Semantic Group

We report Δ Recall@1, defined as the difference between LongBEL and the sentence-level baseline on recurring mentions. Figures S3a and S3b show the results for LongBEL-1B and LongBEL-8B by semantic group and dataset.

The plots show how the gains are distributed across entity types. On MM-ST21pv, most semantic groups have positive Δ Recall@1 values. On QUAERO-EMEA and SPACCC, the estimates are more variable, especially for groups with small support. For this reason, the group-level results should be interpreted together with the number of recurring mentions reported next to each semantic group.

Semantic group	Support	1B		8B	
		Base	SynCABEL	Base	SynCABEL
PROC	586	**74.4**±3.5	73.2±3.5	74.9±3.6	**77.5**±3.5
OBJC	40	**82.5**±12.2	78.3±13.4	**82.5**±12.6	80.0±13.1
PHEN	58	82.8±9.5	**87.0**±8.8	81.0±10.4	**84.5**±8.9
GEOG	51	**98.0**±3.1	96.9±4.9	**98.0**±3.4	96.1±5.0
DEVI	33	**54.5**±17.0	49.3±18.3	54.5±19.4	**69.7**±15.6
LIVB	311	**85.2**±3.9	82.9±4.5	**85.2**±3.9	83.9±4.2
PHYS	161	68.9±6.9	**69.1**±6.8	**75.2**±7.2	75.0±6.9
DISO	1,004	73.4±2.9	**75.2**±2.8	73.8±2.8	**77.6**±2.7
CHEM	340	81.5±4.0	**83.4**±4.0	82.6±3.9	**85.3**±3.8
ANAT	494	75.1±3.9	**78.3**±3.6	79.6±3.4	**80.0**±3.6
Overall	3,078	76.2±1.6	**77.0**±1.6	77.6±1.6	**79.7**±1.6

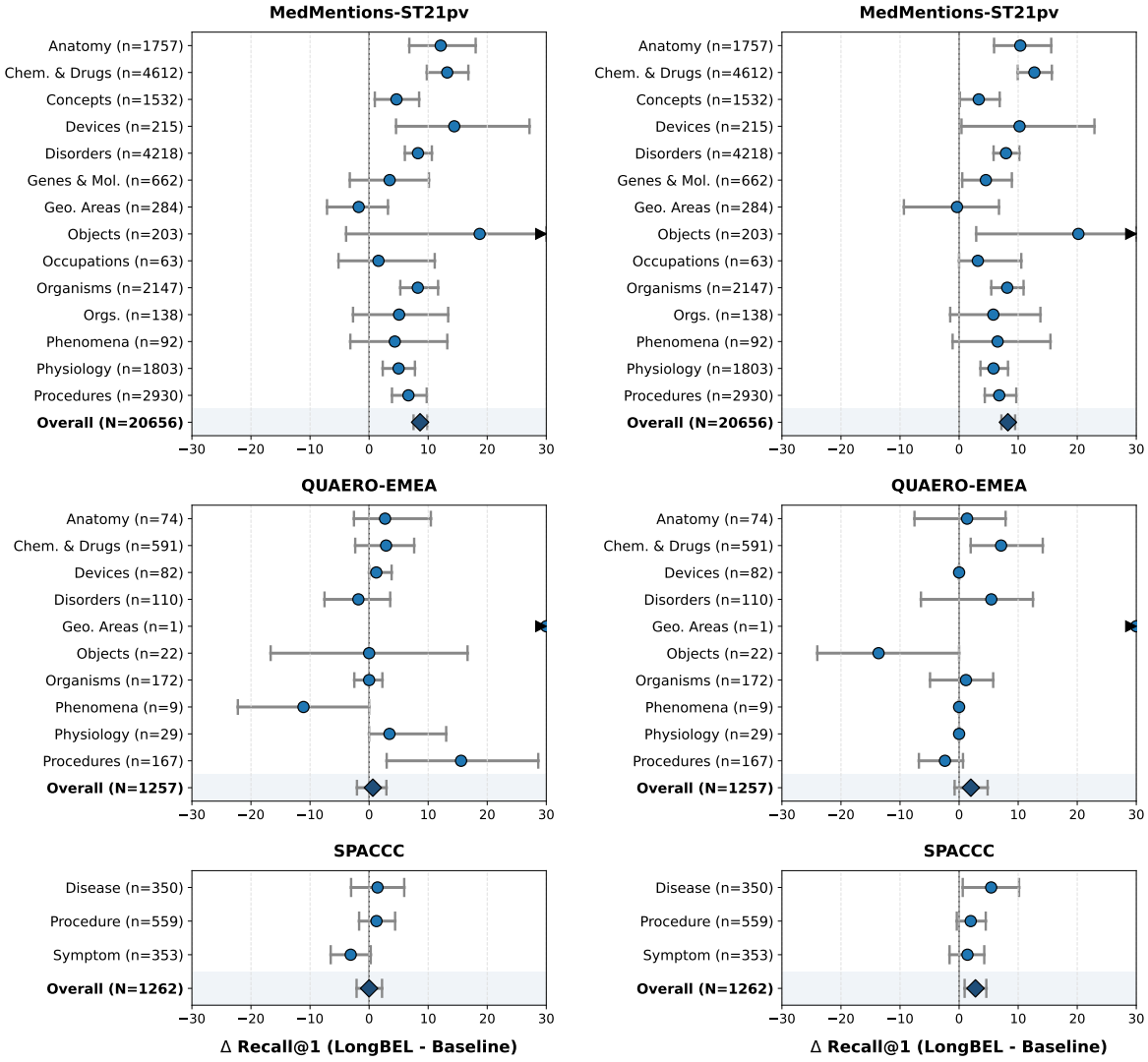
Table S13: Recall@1 by semantic group on MEDLINE. Results are reported with 95% confidence intervals estimated by document-level bootstrap.

Semantic group	Support	1B			8B		
		Base	LongBEL	SynCABEL	Base	LongBEL	SynCABEL
PROC	269	83.3±6.8	81.8±7.2	**84.4**±8.0	85.9±8.1	83.3±7.6	**87.4**±6.2
OBJC	54	64.8±16.8	64.8±21.3	**68.5**±14.7	66.7±19.7	61.1±21.9	**75.9**±13.2
PHEN	36	**77.8**±15.6	75.0±14.6	69.4±15.9	69.4±16.3	69.4±16.3	**75.0**±14.5
GEOG	12	83.3±16.6	**91.7**±9.2	**91.7**±15.0	83.3±16.6	**100.0**±0.0	91.7±9.2
DEVI	99	37.4±21.9	37.4±22.6	**38.4**±22.5	37.4±21.9	**38.4**±22.5	37.4±21.9
LIVB	265	76.2±8.4	76.6±7.9	**80.8**±8.8	77.0±8.8	79.2±9.3	**83.8**±6.6
PHYS	83	77.1±10.2	**78.3**±10.0	77.1±9.2	84.3±8.0	83.1±6.8	**89.3**±8.1
DISO	314	**78.0**±5.6	71.0±6.5	76.8±4.3	73.6±5.6	75.5±7.6	**76.8**±5.9
CHEM	880	73.8±7.0	74.3±6.3	**75.6**±6.5	71.5±5.0	74.5±4.8	**79.0**±7.0
ANAT	184	84.2±7.2	**87.0**±5.6	77.2±8.1	80.4±5.5	**82.6**±6.5	**82.6**±6.1
Overall	2,196	75.1±3.6	74.5±3.7	**75.8**±3.0	73.8±4.0	75.4±3.4	**79.0**±3.2

Table S14: Recall@1 by semantic group on EMEA. Results are reported with 95% confidence intervals estimated by document-level bootstrap.

Semantic group	Support	1B			8B		
		Base	LongBEL	SynCABEL	Base	LongBEL	SynCABEL
SYMP	2,898	60.5±2.5	59.8±2.5	**61.3**±2.7	61.7±2.5	62.0±2.6	**63.8**±2.5
PROC	3,582	67.4±2.1	66.6±2.1	**68.0**±2.3	69.2±2.1	69.0±2.1	**70.1**±2.1
DISO	2,341	62.5±2.3	61.9±2.4	**65.0**±2.4	63.2±2.5	63.6±2.1	**66.3**±2.2
Overall	8,821	63.8±1.6	63.1±1.6	**65.0**±1.6	65.1±1.5	65.3±1.5	**67.0**±1.5

Table S15: Recall@1 by semantic group on SPACCC. Results are reported with 95% confidence intervals estimated by document-level bootstrap.



(a) LongBEL-1B.

(b) LongBEL-8B.

Figure S3: Difference in Recall@1 between LongBEL and the sentence-level baseline, computed on mentions of repeated concepts only. Results are reported per semantic group across datasets for the 1B and 8B models.

C.3 Robustness to Noisy Memory

To evaluate whether a model propagates errors through memory, we define the *Copy Wrong Memory Error Rate* (CWME). CWME measures how often a model repeats an earlier incorrect prediction for the same ground-truth concept.

Let $M_e = (m_1, m_2, \ldots, m_k)$ be the ordered mentions of concept e in a document. Suppose the model first predicts an incorrect concept $e' \neq e$ at mention m_f. All later mentions of e are then *exposed mentions*, since the wrong prediction is present in memory:

$$E_e = \{m_i \in M_e \mid i > f\}.$$

An *echoed error* occurs when the model repeats the same wrong prediction e':

$$R_e = \{m_i \in E_e \mid \hat{e}_i = e'\}.$$

CWME is computed as:

$$\text{CWME} = \frac{\sum_{e \in \mathcal{E}} |R_e|}{\sum_{e \in \mathcal{E}} |E_e|} \times 100,$$

where $\mathcal{E}$ includes concepts with an initial error and at least one later exposed mention. Lower CWME indicates fewer cascading errors. Figures S4a and S4b compare gold-memory training with robust cross-validated memory training using

$$\Delta\text{CWME} = \text{CWME}_{\text{robust memory}} - \text{CWME}_{\text{gold memory}}.$$

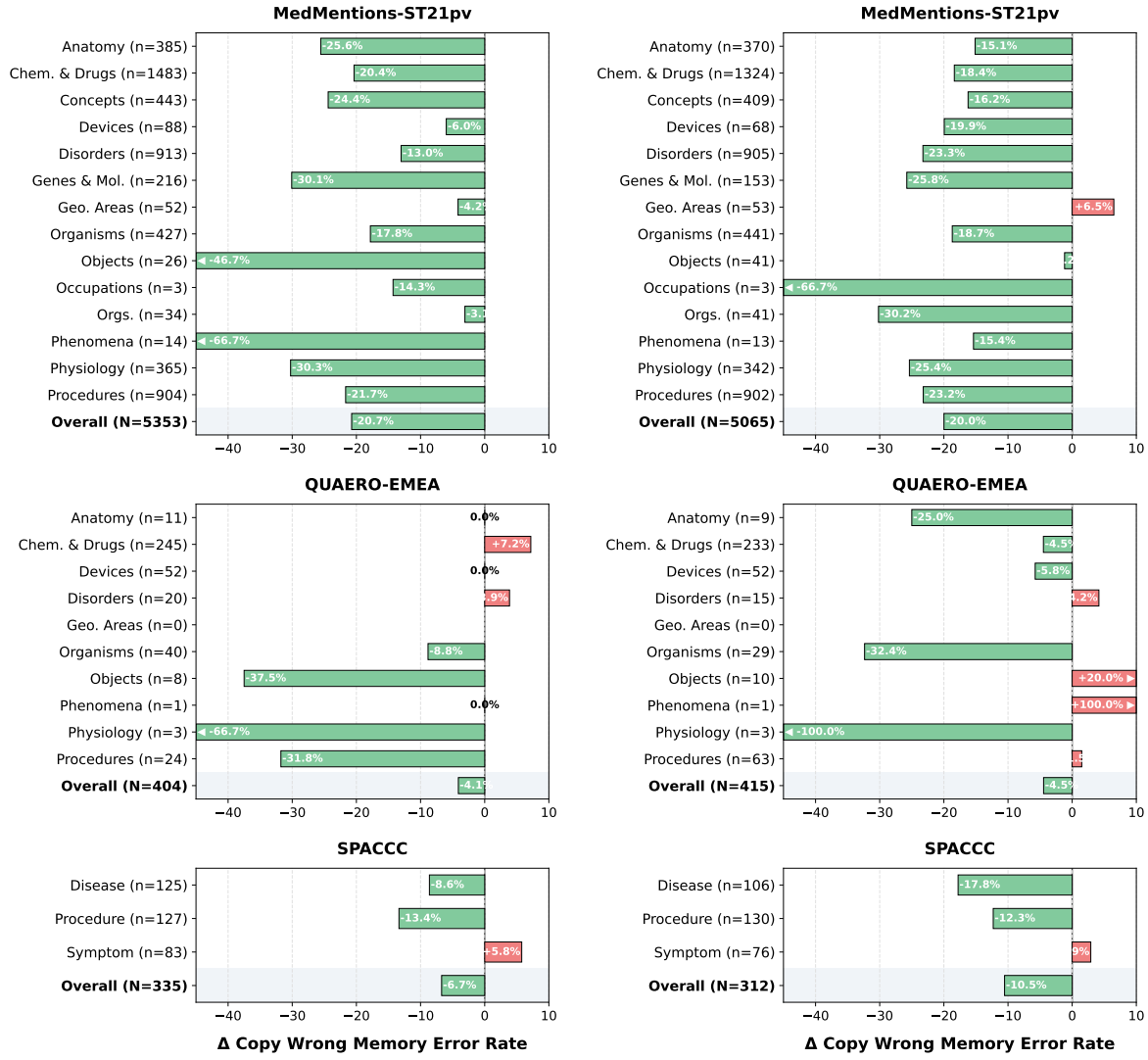
Negative values indicate fewer copied errors with robust memory.

Overall, robust memory consistently reduces copied errors. For LongBEL-1B, CWME decreases by 20.7 points on MM-ST21pv, 4.1 points on QUAERO-EMEA, and 6.7 points on SPACCC. For LongBEL-8B, the corresponding reductions are 20.0, 4.5, and 10.5 points, respectively. These results show that training with predicted memory makes the model less likely to repeat earlier mistakes. The reduction is strongest and most stable on MM-ST21pv, where many semantic groups contain enough exposed mentions to provide reliable estimates. In contrast, some groups in QUAERO-EMEA and SPACCC contain very few examples, leading to noisier category-level results and occasional positive deltas. This is consistent with our main analysis: robust memory is most effective when enough recurring concepts are available to learn reliable consistency patterns.

C.4 End-to-End Threshold Selection Curves

Figure S5 shows how end-to-end precision, recall and F1-score vary with the normalization confidence threshold. A normalized prediction is accepted only if its score is above this threshold; otherwise, it is rejected. The threshold therefore controls the trade-off between precision and recall.

The score has a different meaning depending on the normalization model. For CODER-all and SapBERT-all, it is the cosine similarity between the mention embedding and the candidate concept embedding. For SynCABEL and LongBEL, it is the confidence score associated with the generated concept. Thresholds are therefore selected separately for each dataset and each model.



(a) LongBEL-1B.

(b) LongBEL-8B.

Figure S4: Change in Copy Wrong Memory Error Rate when using robust memory instead of gold memory. Negative values indicate fewer cascading errors. Results are computed on exposed recurring concepts and reported by semantic group for the 1B and 8B models.

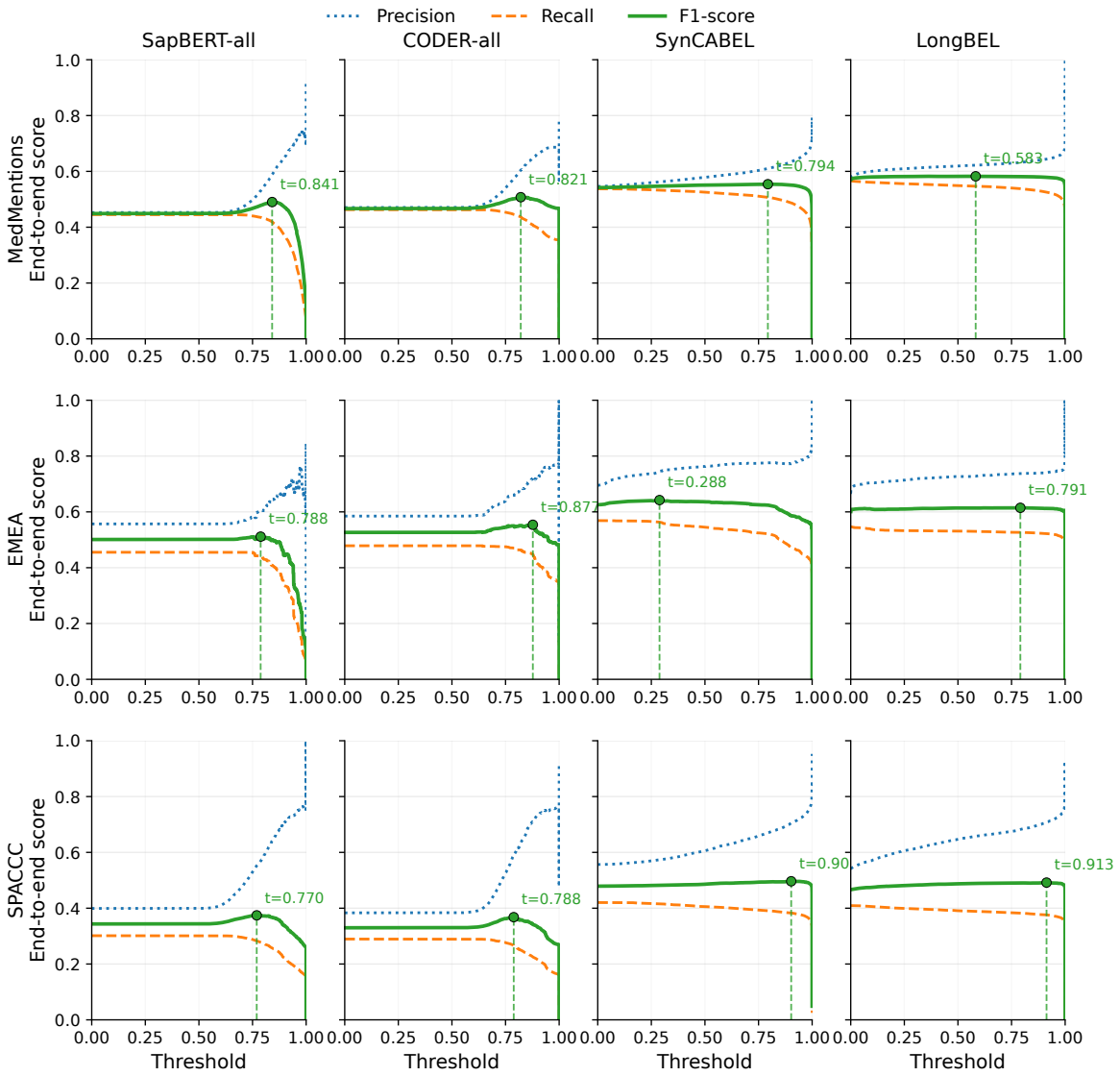


Figure S5: End-to-end precision, recall and F1-score as a function of the normalization confidence threshold. Rows correspond to datasets, ordered as MedMentions, QUAERO-EMEA and SPACCC. Columns correspond to normalization models. Dashed vertical lines indicate the selected threshold for each model.

For each dataset and each normalization model, the threshold is chosen on the validation set by maximizing end-to-end F1-score over 1,000 candidate thresholds sampled from the empirical score distribution. The selected threshold is then fixed and applied to the test set for the final results reported in Table 4.12.

List of Figures

1 Iceberg representation of EHR data: structured data form the visible and easily accessible portion, whereas unstructured data form the larger, submerged portion, which contains richer information but is more difficult to access. 2

1.1 Simplified overview of an EHR-based hospital information system. A non-exhaustive set of care activities produces data recorded in the EHR, which can then be transferred to a clinical data warehouse (CDW) for secondary use. Inspired by Laforet (2017). 8

1.2 Example of a clinical data warehouse pipeline. The Stanford STARR ecosystem illustrates the transformation of heterogeneous patient data sources into research datasets, analytical tools, and software within a secure data science infrastructure. Although this example comes from Stanford, the same conceptual steps are relevant to other large hospital CDWs, including the AP-HP CDW. Reproduced from Callahan et al. (2023). 10

1.3 Progressive EHR digitization over time in the AP-HP clinical data warehouse. The figure shows the normalized number of data points available for several EHR functionalities between 2013 and 2022, using data from the AP-HP clinical data warehouse. Different functionalities become available at different periods, illustrating why calendar time in a CDW is also a data-availability variable. Reproduced from the `EDS-TeVa` work (Remaki et al., 2023b). 12

1.4 Example of biomedical named entity recognition on a synthetic clinical text. Relevant spans are identified and assigned semantic categories, such as diseases, symptoms, drugs, laboratory tests, or demographic information. The example illustrates that named entity recognition requires both selecting the correct span boundaries and assigning the correct entity type. 16

1.5 Comparison of common formulations for named entity recognition. In BIO tagging, each token receives a label indicating whether it begins an entity, continues an entity, or lies outside any entity. BIOUL adds more explicit boundary information by distinguishing beginning, inside, last, unit-length, and outside tokens. In span prediction, the system directly detects and classifies complete text spans instead of reconstructing entities from token-level labels. 17

1.6 Illustration of clinical entity qualification. BioNER detects clinical mentions, while qualification assigns contextual attributes that determine whether the mention is negated, hypothetical, historical, or related to family history. 21

1.7 Example of biomedical entity normalization. After entity mentions have been detected in the clinical text, each mention is mapped to its corresponding standardized concept code in a KB and concept name. This step makes different textual forms computable as the same biomedical concept. 23

2.1 Overview of the end-to-end patient phenotyping pipeline. Structured and unstructured data are extracted from electronic health records, enabling large information retrieval, refining cohort selection, and facilitating more robust patient comparisons. ATC: Anatomical Therapeutic Chemical; SNOMED: Systematized Nomenclature of Medicine Clinical Terms. 31

2.2 Cohort selection flowchart. Starting with integrating biology and bedside extraction of 1,947,870 stays (38,384 patients), the cohort was filtered to include stays with at least 1 International Classification of Diseases, 10th Revision (ICD-10) code corresponding to the studied diseases and at least 1 recorded hospital discharge summary. A final study cohort of 18,604 stays (6891 patients) was created, with 103 randomly selected discharge summaries annotated for training and validation purposes. NLP: natural language processing. 36

2.3 Overview of the two NLP pipelines applied to discharge summaries. The laboratory test pipeline extracts and normalizes laboratory tests and their measurements, while the drug pipeline extracts, qualifies, and normalizes drug treatments. 38

2.4 Diagram of the named entity recognition and qualification architecture for laboratory test and drug entities. BERT: bidirectional encoder representations from transformers; CNN: convolutional neural network; CRF: conditional random field. 39

2.5 Diagram of the laboratory test measurement extraction process. HBV: hepatitis B virus; HCV: hepatitis C virus. 41

2.6 Diagram of the normalization process. In this example, the best match of “paracetomol,” written with a typo, is “paracetamol,” which provides the correct Anatomical Therapeutic Chemical (ATC) code. 42

2.7 Age distributions and hospitalizations over time in the study cohort. The left column shows the age distribution of patients in five-year intervals. The right column shows the number of hospitalizations over time from 2012 to 2021 in quarterly intervals. 43

2.8 Evolution of NER performance according to the number of annotated APHP discharge summaries used for training. The curve shows the F1-score, with the shaded area representing the 95% confidence interval. 46

2.9 Distribution of selected quantitative laboratory test results in the study cohort. For each laboratory test, the first row shows values available from structured data only, the second row shows the additional values extracted from unstructured discharge summaries, and the third row shows the combination of structured and unstructured data. The analysis was performed on the whole study cohort, pooling the 4 studied diseases. CRP: C-reactive protein; INR: international normalized ratio; GFR: glomerular filtration rate. 52

3.1 Example of contextual ambiguity in biomedical entity linking. The same abbreviation, *AS*, can refer to different biomedical concepts depending on the surrounding context, such as aortic valve stenosis or ankylosing spondylitis. 58

3.2 Illustration of the motivation behind SynCABEL. Human-annotated data provide contextual examples but limited concept coverage, while knowledge bases provide broad coverage but no direct contextual supervision. SynCABEL generates contextualized examples from knowledge-base concepts to increase the coverage of context-aware biomedical entity linking models. 59

3.3 Example of the disambiguation process used to select a natural representation e_m = "Fluid Discharge" for a concept e = "C0012621" and a mention m = "discharge". 62

3.4 Overview of SynCABEL. LLM: large language model 63

3.5 LLM-as-a-judge evaluation of predicted vs. gold concepts using four clinical-relation classes: Correct, Broad, Narrow, and No relation. 64

3.6 Data-efficiency analysis. Exact-match Recall@1 performance as a function of human training data availability. The dashed line marks the performance of the standard baseline trained on 100% of the human-annotated data. Stars ($\star$) indicate where the SynCABEL-augmented model reaches or exceeds this baseline. The 0% setting corresponds to training on synthetic data only. Error bars denote 95% bootstrap confidence intervals. . . 73

4.1 Illustrative example of document-level context and memory. **Top:** Restricted to local context, a standard model mislinks *[PET]* to *Petechial haemorrhage*. **Bottom:** By using global clinical context and previous predictions, LongBEL correctly links *[PET]* to *Pre-eclamptic toxemia*. 82

4.2 Illustration of robust memory construction with out-of-fold predictions. The training set is split into five folds. For each fold, a base model is trained on the remaining folds and predicts the held-out fold. These held-out predictions are then used to create the memory for final LongBEL training, so that the memory used during training better matches inference-time conditions. 86

4.3 Overview of LongBEL inference. For each mention, LongBEL combines local context, full-document context, and memory of previous predictions to generate multiple ranked candidate lists using semantic-guided constrained decoding. These lists are merged with Reciprocal Rank Fusion (RRF) to produce the final prediction. The predicted concept is then added to memory and used for subsequent mentions. 87

4.4 Saliency maps for LongBEL-8B predictions. Failure case from MedMentions-ST21pv. In the IFN-γ example, LongBEL-8B gives high saliency to an incorrect previous prediction, showing that an early linking error can influence a later abbreviation prediction through memory. 100

4.5 Saliency maps for LongBEL-8B predictions. These two examples illustrate cases where document-level evidence helps. In the ND example, LongBEL gives high saliency to the earlier definition "neurological deterioration (ND)" and to the corresponding memory entries. In the PET example, it uses obstetric context and previous linked mentions, especially hypertension and proteinuria, to link PET to pre-eclamptic toxemia. 101

5.1 Conceptual answer to the iceberg representation introduced in Figure 1. After NLP, part of the information contained in clinical notes can be extracted, qualified, and normalized into structured variables. The original notes remain below the surface, emphasizing that NLP enriches structured EHR data without replacing clinical narratives. 106

5.2 Proof-of-concept tool for similar-patient retrieval from discharge summaries. (A) Prototype interface used to input a target discharge summary, select the body systems to compare, and set the number of similar discharge summaries to retrieve. (B) Fictitious example of a target discharge summary after extraction of clinical information by the pipeline. (C) Aggregate summary of treatments and outcomes among fictitious retrieved patients. All patient information shown in this figure is fictitious and used only for illustration. This prototype demonstrates technical feasibility and is now being further developed and evaluated as part of ongoing work. 110

6.1 Représentation en iceberg des données issues du dossier patient informatisé (DPI) : les données structurées constituent la partie émergée, directement accessible et exploitable, tandis que les données non structurées représentent une part plus riche en informations, mais plus difficile à exploiter. 116

6.2 Vue d'ensemble de la chaîne de traitement pour l'enrichissement du phénotypage. 118

6.3 Chaîne de traitement des examens biologiques à partir des comptes rendus d'hospitalisation. 119

6.4 Chaîne de traitement des traitements médicamenteux à partir des comptes rendus d'hospitalisation. 119

6.5 Vue d'ensemble de SynCABEL : génération d'exemples contextualisés à partir des concepts d'une base de connaissances, combinaison avec les données annotées et normalisation guidée vers les concepts valides. 123

6.6 Vue d'ensemble de LongBEL : prise en compte du contexte local, du document complet et des prédictions précédentes pour la normalisation séquentielle des mentions, suivie d'une fusion des prédictions par Reciprocal Rank Fusion. 126
6.7 De la donnée non structurée à la donnée structurée : le TAL permet d'extraire, qualifier et normaliser une partie des informations contenues dans les notes cliniques. 129
S1 End-to-end precision, recall and F1-score as a function of the normalization confidence threshold. Rows correspond to datasets, ordered as MedMentions, QUAERO-MEDLINE, QUAERO-EMEA and SPACCC. Columns correspond to normalization models. Dashed vertical lines indicate the selected threshold for each model. 162
S2 Distribution of MMD^2 values comparing human-annotated concepts with semantic-group-stratified random samples, and two independent stratified random samples. Values are shown on a logarithmic scale. 163
S3 Difference in Recall@1 between LongBEL and the sentence-level baseline, computed on mentions of repeated concepts only. Results are reported per semantic group across datasets for the 1B and 8B models. 166
S4 Change in Copy Wrong Memory Error Rate when using robust memory instead of gold memory. Negative values indicate fewer cascading errors. Results are computed on exposed recurring concepts and reported by semantic group for the 1B and 8B models. 168
S5 End-to-end precision, recall and F1-score as a function of the normalization confidence threshold. Rows correspond to datasets, ordered as MedMentions, QUAERO-EMEA and SPACCC. Columns correspond to normalization models. Dashed vertical lines indicate the selected threshold for each model. 169

List of Tables

1.1 Synthesis of BioNER model families according to their strengths and limitations . . . 19
1.2 Synthesis of clinical entity qualification model families according to their strengths and limitations . . . 21
1.3 Synthesis of BEL model families according to their strengths and limitations 26

2.1 Description of the knowledge bases used for laboratory tests and drug treatments. . . . 37
2.2 Architecture and fine-tuning configuration of the final supervised NER and qualification models. CNN: convolutional neural network; CRF: conditional random field. . . . 40
2.3 Data description of the study cohort. Number of patients, hospitalizations, and discharge summaries for each studied disease. The number of discharge summaries is higher than the number of hospitalizations, as patients may change departments several times during the same stay (e.g., transfer to an intensive care unit). . . . 43
2.4 Performance of the rule-based NER approach on the 20 AP-HP discharge summaries of the test set. The approach was evaluated for drug names and laboratory test names, corresponding to the entity types covered by the terminology dictionaries. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals. . . . 44
2.5 Comparison of pretrained encoders for supervised NER on the AP-HP test set. All models were evaluated using the same NER configuration and the same 20 held-out discharge summaries. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals. . . . 45
2.6 Ablation study of the CNN and CRF components for NER using CamemBERT-EDS. All configurations were evaluated on the same 20 AP-HP discharge summaries. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals. . . . 45

2.7 Performance of the final CamemBERT-EDS–CNN–CRF model for NER on the 20 AP-HP discharge summaries of the test set. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals. 45
2.8 Comparison of supervised, open-vocabulary, and instruction-tuned NER approaches on the QUAERO EMEA and MEDLINE corpora. Precision, recall, and F1-score were computed using strict exact matching and are reported with document-level bootstrap 95% confidence intervals. Few-shot LLM demonstrations were retrieved according to TF–IDF similarity with the input document. Qwen3-4B-Instruct was run locally, whereas GPT-5.6 Luna was accessed through the OpenAI API. Inference time is reported in seconds per document; API-based timings include communication and network latency and are therefore not directly comparable to local inference times. 47
2.9 Ablation study of the CNN component for the qualification task on the AP-HP discharge summaries. Both configurations were evaluated on the same 20 held-out discharge summaries. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals. 48
2.10 Performance of the final qualification model on the 20 held-out AP-HP discharge summaries. Precision, recall, and F1-score are reported for non-default qualification values, with document-level bootstrap 95% confidence intervals. 48
2.11 Performance of the model for measurement extraction. This rule-based algorithm was evaluated on 20 annotated discharge summaries. Each result was bootstrapped by discharge summary to provide a 95% CI. 48
2.12 Comparison of normalization methods on 11 annotated AP-HP discharge summaries. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals. 49
2.13 Performance of the selected normalization methods on 11 annotated AP-HP discharge summaries and on the QUAERO FrenchMed corpus. CODER-all was used for laboratory test normalization and Jaro–Winkler distance for drug normalization. The QUAERO evaluation included 3 EMEA documents and 833 MEDLINE titles. Precision, recall, and F1-score are reported with document-level bootstrap 95% confidence intervals. 49
2.14 Performance of the models for the combined named entity recognition and normalization tasks on AP-HP CDW discharge summaries. The model was tested on 11 discharge summaries. Each result was bootstrapped by discharge summary to provide a 95% confidence interval (CI), shown in parentheses. 50
2.15 Number of patients with positive antibodies identified from structured data and from the combination of structured and unstructured data. The last column reports the additional number of patients identified only through unstructured data. 51
2.16 Number of patients with drug treatments identified from structured data and from the combination of structured and unstructured data. The last column reports the additional number of patients identified only through unstructured data. 53

3.1 Summary statistics of the human-annotated evaluation datasets. Average length: average document length in characters; Clinical cases: whether the dataset consists of clinical case reports; # Sem. Groups: number of UMLS semantic groups represented in the dataset; Coverage: proportion of concepts from the target terminology represented in the dataset. MM-ST21pv: MedMentions-ST21pv; SPACCC: Spanish Clinical Case Corpus. 65
3.2 Summary statistics of the knowledge base subsets used for candidate retrieval across the evaluated datasets. 65
3.3 Summary of synthetic datasets. 66
3.4 Examples of synthetic contextualized sentences generated by SynCABEL for the three target datasets. The concept column reports the concept identifier and title, and brackets indicate the generated mention used for entity linking. 66
3.5 Supervised fine-tuning hyperparameters used across the evaluated generative models. 67
3.6 Inference hyperparameters used during concept generation. 67
3.7 Entity linking performance (Recall@1) on biomedical benchmarks. The highlighted row shows our contribution, where the base model is trained on an augmented dataset. The best results are shown in **bold**, the second-best results are underlined, and the "Average" column reports the mean score across the four benchmarks. 95% confidence intervals are estimated using empirical bootstrap resampling at the document level. 68
3.8 Recall@1 scores for different concept representation methods. mBART-large was supervised only on the human-annotated dataset (i.e., without data augmentation), and inference was performed using the guided inference method. The highlighted row shows the chosen representation method. 69
3.9 Recall@1 scores for different training data composition strategies. The highlighted row shows the chosen strategy. 69
3.10 Distribution of clinician labels among code mismatches in the manually annotated subset of SPACCC. Percentages are computed within each semantic group, and overall percentages are computed over the full annotated subset. 70
3.11 Llama-3-8B performance on SPACCC with and without SynCABEL augmentation. Results are reported as percentages with 95% confidence intervals estimated via document-level empirical bootstrap, for exact code matching and each LLM-as-a-judge label. The Δ score column reports the absolute difference between the two settings in percentage points. 71
3.12 Exact-match Recall@1 performance on four biomedical benchmarks comparing two variants of Llama-3-8B: one trained only on the corresponding human-annotated dataset with guided inference, and one further enhanced with SynCABEL. Results are reported overall and for subsets defined by seen/unseen concepts. The Δ score column reports the absolute difference between the two settings, computed as w/ SynCABEL minus w/o SynCABEL on the reported score scale. 95% confidence intervals were estimated via the empirical bootstrap method at the document level. . . . 72

3.13 Inference speed and memory footprint. Model size refers to the memory required by model parameters, candidate size refers to the memory required by the candidate representation, and speed is reported in mentions per second. 74

3.14 End-to-end named entity recognition (NER) and normalization performance. For each dataset, the *NER only* row reports mention detection performance before normalization. The following rows add a normalization model on top of the same NER predictions. A prediction is considered correct only when both the entity mention and the normalized concept are correctly identified. Nested entities with the same label were filtered before evaluation by retaining only the outermost span. 75

4.1 Summary statistics of the evaluation datasets used in this chapter. Average length is reported in characters. Coverage is the proportion of concepts from the target terminology represented in the dataset. % First is the percentage of mentions whose concept appears for the first time in a document. % Recurring is the percentage of mentions whose concept has already appeared at least once earlier in the same document. MM-ST21pv: MedMentions-ST21pv; SPACCC: Spanish Clinical Case Corpus. 88

4.2 Entity linking performance (Recall@1) on biomedical benchmarks. The highlighted rows correspond to the LongBEL models introduced in this chapter. The best results are shown in **bold**, the second-best results are underlined, and the "Average" column reports the mean score across the three benchmarks. 95% confidence intervals are estimated using empirical bootstrap resampling at the document level. 89

4.3 Ablation study (recall) evaluating LongBEL-1B and LongBEL-8B. 'Ctx' denotes the inclusion of global context, and 'Mem.' indicates whether memory is enabled (✓) or disabled (✗). Best results are in **bold**, second-best are underlined. 95% confidence intervals are estimated using empirical bootstrap resampling at the document level. The ensemble rows fuse all components via RRF. 90

4.4 Recall comparison across datasets, broken down by concept frequency within documents for both the 1B and 8B models. "Baseline" corresponds to the underlying Llama models applied at the sentence level without cross-mention memory. "LongBEL" extends this model with global document context and robust prediction memory. "First" denotes mentions whose concept appears for the first time in the document, while "Recurring" denotes mentions whose concept has already appeared earlier in the document. 95% confidence intervals were estimated via the empirical bootstrap method at the document level. 91

4.5 Detailed subset results on MM-ST21pv. The first two rows report overall Recall@1 and Recall@5; all other rows report Recall@1 by subset. Support gives the percentage and number of examples. Scores are reported with half of the 95% confidence interval. Base denotes the baseline model and Ens. the ensemble. 93

4.6 Detailed subset results on QUAERO-EMEA. The first two rows report overall Recall@1 and Recall@5; all other rows report Recall@1 by subset. Support gives the percentage and number of examples. Scores are reported with half of the 95% confidence interval. Base denotes the baseline model and Ens. the ensemble. 94
4.7 Detailed subset results on SPACCC. The first two rows report overall Recall@1 and Recall@5; all other rows report Recall@1 by subset. Support gives the percentage and number of examples. Scores are reported with half of the 95% confidence interval. Base denotes the baseline model and Ens. the ensemble. Simple and composite concept rows are reported only for SPACCC. 95
4.8 Qualitative examples where document-level context helps LongBEL-8B. The first example shows an explicit abbreviation definition, while the second shows a case requiring broader clinical context. The ND example is from MedMentions-ST21pv; the PET example is manually constructed. . 96
4.9 Qualitative examples where document-level memory does not improve over the baseline. The examples illustrate error propagation, failure on a specialized chemical abbreviation, and a simple case where both models succeed. All examples are from MedMentions-ST21pv. 97
4.10 Cross-dataset biomedical entity linking performance in Recall@1 (%). Each model is trained on the dataset indicated in the rows and evaluated, without further training, on the dataset indicated in the columns. All models use the same Llama-3-8B backbone. Values are reported as the point estimate $\pm$ half of the 95% confidence interval. The best score for each training–evaluation dataset pair is shown in bold. Dashes indicate unavailable configurations. 99
4.11 Inference speed and memory footprint. Model size refers to the memory required by model parameters, candidate size refers to the memory required by the candidate representation, and speed is reported in mentions per second. 102
4.12 End-to-end named entity recognition (NER) and normalization performance. For each dataset, the *NER only* row reports mention detection performance before normalization. The following rows add a normalization model on top of the same NER predictions. A prediction is considered correct only when both the entity mention and the normalized concept are correctly identified. 103
4.13 Performances en Recall@1 des systèmes de normalisation d'entités biomédicales sur les quatre jeux de données évalués. Les lignes « + SynCABEL » correspondent aux modèles entraînés avec augmentation par données synthétiques. 124
4.14 Performances en Recall@1 des systèmes de normalisation d'entités biomédicales sur les trois jeux de données évalués. Les lignes en surbrillance correspondent aux modèles LongBEL introduits dans ce chapitre. Les meilleures performances pour chaque jeu de données et pour la moyenne sont indiquées en gras. 127
S1 Queries used to extract the project data from the CDW research database. 147

S2 Descriptive labels provided as input to the GLiNER models for the QUAERO FrenchMed NER experiments. French labels were used with the French-language model and English labels with the English-language models. . . 150
S3 Performance of the NER algorithms built upon DrBert (Labrak et al., 2023) or CamemBert-EDS (Dura et al., 2022) evaluated on AP-HP discharge summaries. The methods were tested on 20 discharge summaries. Each result was bootstrapped by discharge summary to provide a 95% confidence interval given inside the parentheses. Best scores in bold. 152
S4 Performance of the NER algorithms built upon CamemBert-Base (Martin et al., 2020) or CamemBert-Bio (Touchent et al., 2024) evaluated on AP-HP discharge summaries. The methods were tested on 20 discharge summaries. Each result was bootstrapped by discharge summary to provide a 95% confidence interval given inside the parentheses. Best scores in bold. 153
S5 Performance of the model for named entity recognition on the MEDLINE corpus. The model was trained on 1664 annotated titles and tested on 833 titles. Each result was bootstrapped by title to provide a 95% confidence interval given inside the parentheses. 153
S6 Performance of the model for named entity recognition on the EMEA corpus. The model was trained on 6 annotated documents and tested on 3 documents. Each result was bootstrapped by document to provide a 95% confidence interval given inside the parentheses. 154
S7 List of Concept Unique Identifiers (CUIs) of the Unified Medical Language System (UMLS), restricted to the Laboratory Procedure semantic type and SNOMED CT vocabulary, for the studied autoantibodies and laboratory tests. 154
S8 List of codes from the Anatomical Therapeutic Chemical (ATC) classification system for the studied drug treatments. 155
S9 Preliminary comparison of encoder–decoder backbones for biomedical entity linking. The table reports Recall@1 on MM-ST21pv, QUAERO-MEDLINE, and QUAERO-EMEA. For each backbone, we compare standard fine-tuning, guided inference, and SynCABEL-augmented training. The values in parentheses indicate the absolute difference between SynCABEL and guided inference for the same backbone. The highlighted rows correspond to models trained with SynCABEL. The best results are shown in **bold**, and the second-best results are underlined. 156
S10 Named entity recognition performance on MM-ST21pv. Confidence intervals are computed by document-level bootstrap and reported at 95%. . . 160
S11 Named entity recognition performance on SPACCC. Confidence intervals are computed by document-level bootstrap and reported at 95%. 160
S12 Recall@1 by semantic group on MedMentions. Results are reported with 95% confidence intervals estimated by document-level bootstrap. 164
S13 Recall@1 by semantic group on MEDLINE. Results are reported with 95% confidence intervals estimated by document-level bootstrap. 165
S14 Recall@1 by semantic group on EMEA. Results are reported with 95% confidence intervals estimated by document-level bootstrap. 165

S15 Recall@1 by semantic group on SPACCC. Results are reported with 95% confidence intervals estimated by document-level bootstrap. 165