

Classification multi-label de données médicales par LSTM temporel et clustering flou

Abdelhamid Gaddari^{*,**}, Haytham Elghazel^{*}, Rakia Jaziri^{***}, Mohand Hacid-Saïd^{*}
Pierre-Henri Comble^{**}

* Université Lyon 1, LIRIS, UMR CNRS 5205, F-69622

prenom.nom@liris.cnrs.fr

** CEGEDIM Insurance Solutions, France

prenom.nom@cegedim.com

*** Université Paris 8, Laboratoire Paragraphe, EA 349

rjaziri@univ-paris8.fr

Résumé. La prévention de la santé est un aspect très important de la recherche en informatique médicale grâce à la prédiction d'événements médicaux. Dans ce travail, nous proposons une approche d'apprentissage profond pour effectuer des prédictions multi-label sur les actes de soins médicaux. L'approche proposée utilise un réseau LSTM temporel (*time-aware long short-term memory*) et l'étend avec des informations supplémentaires à partir d'un clustering flou du même portefeuille. Le premier mécanisme (*time-aware*) est utilisé pour gérer l'irrégularité temporelle entre les éléments d'une trajectoire médicale, tandis que le second mécanisme (*fuzzy clustering*) aide à modéliser l'hétérogénéité entre les patients et les traitements. En utilisant un large portefeuille d'actes médicaux remboursés (plus de 16 millions d'actes médicaux consommés) par une assurance médicale en France, nous montrons que notre approche surpasse les méthodes traditionnelles et d'apprentissage profond dans la classification médicale multi-label. Notre travail a pour objectif de soutenir la prévention médicale et plus largement améliorer la qualité des services de santé et des assurances.

1 Introduction

La recherche en informatique médicale a connu un développement rapide au cours de ces dernières années, et plus particulièrement l'utilisation de modèles d'apprentissage automatique pour différentes tâches médicales (Hong et al., 2018), telles que l'évolution de maladies ou la prédiction de diagnostics. Dans ce papier, nous sommes confrontés à un sujet commun de la recherche biomédicale : la prédiction multi-label des actes de soins médicaux. Nous utilisons les données d'une assurance médicale privée en France, représentant tous les actes de soins médicaux consommés par les assurés. Cependant, nous devons faire face au défi de l'irrégularité temporelle qui réside dans les données de santé, puisque les intervalles entre les différents actes de soins varient de quelques jours à plusieurs mois et dépendent de la pathologie médicale concernée. Un autre défi motivant notre travail est l'hétérogénéité que présentent nos données, puisque le portefeuille est composé d'assurés de tous les âges et avec des situations médicales

et socio-démographiques différentes (*inter-hétérogénéité*). Cela engendre de multiples profils de soins, en termes de longueur des dossiers médicaux, de fréquence et d'intervalles de temps entre les différents actes de soins. En outre, les dossiers d'un même assuré sont eux aussi hétérogènes (*intra-hétérogénéité*), puisqu'un patient peut avoir des situations de soins différentes avec des durées, des gravités, des traitements, et des besoins différents.

Les réseaux de neurones récurrents (RNN), et plus particulièrement leur variante Long Short-Term Memory (LSTM), sont les méthodes d'apprentissage profond les plus prometteuses lorsqu'il s'agit de gérer des données séquentielles de longueur variable. Ils sont également adaptés à la classification multi-label dans l'informatique de santé. Cependant, les réseaux LSTM de base souffrent du problème d'irrégularité temporelle, puisqu'ils supposent que les enregistrements séquentiels ont un intervalle de temps constant entre ses éléments. Il convient également de traiter l'hétérogénéité du portefeuille et du dossier médical (hétérogénéité intra et inter patient), afin d'aider les réseaux LSTM à capter l'interdépendance entre les différents actes de soins.

Pour résoudre ces problèmes, nous proposons une approche qui utilise un réseau LSTM et nous l'étendons avec deux mécanismes (i.e., temporel et clustering flou). Nous procédons d'abord à un clustering flou du portefeuille, définissant des clusters de pathologies homogènes, pour lesquels chaque assuré a des différentes probabilités d'appartenance. Nous utiliserons ensuite ces probabilités comme métadonnées pour renforcer un modèle d'apprentissage profond basé sur le LSTM temporel (Men et al., 2021) qui prend en compte les intervalles de temps entre les éléments du dossier d'un patient. L'objectif est de capturer efficacement les corrélations dans l'espace des variables en entrée afin de faciliter le processus de prédiction multi-label.

2 État de l'art : Modèles séquentiels pour l'apprentissage des données médicales

Les avancées technologiques autour du stockage massif de données et des techniques informatiques ont beaucoup stimulé l'utilisation de techniques séquentielles basées sur les données, pour la prédiction des risques médicaux, le diagnostic ou l'évolution des maladies. Les soins médicaux étant par nature temporels et épisodiques, la modélisation de l'ensemble de la trajectoire médicale est indispensable pour comprendre les relations sous-jacentes entre les différents épisodes, voire au sein d'un même épisode. Des connaissances spécifiques à un domaine ou des méthodes statistiques telles que des modèles de séries chronologiques ou des approches bayésiennes (Orphanou et al., 2014) ont été utilisées afin d'étudier la progression ou le diagnostic de la maladie grâce à la modélisation des parcours de soins.

Ces dernières années, les techniques d'apprentissage profond sont devenues plus populaires dans de nombreux domaines tel que le traitement du langage naturel (Wen et al., 2015). Le succès de l'apprentissage profond a attiré les chercheurs en santé en raison de sa capacité à modéliser des relations non linéaires de longue dépendance entre les éléments des données séquentielles. Par exemple, (Choi et al., 2016) a utilisé un réseau de neurones récurrent (RNN) pour prédire les événements cliniques à partir de données séquentielles horodatées et (Díez-Sanmartín et Cabezuelo, 2020) pour prédire les taux de survie après une greffe de rein. Cependant, dans le cas de séquences plus longues, les RNN souffrent de disparition ou explo-

sion du gradient. Le réseau LSTM (Long Short-Term Memory), introduit pour la première fois par (Hochreiter et Schmidhuber, 1997), est l'une des variantes les plus populaires des RNN abordant ce problème via une architecture multicouche (Cortez et al., 2018). Par ailleurs, il a été prouvé que les RNN et leurs variantes (LSTM, GRU) sont facilement adaptables à la classification multi-label dans le domaine médical (Nam et al., 2017). Néanmoins, ces approches devraient faire face au défi de l'irrégularité temporelle entre les éléments du dossier médical. Motivé par ce constat, (Zhu et al., 2017) ont introduit un modèle LSTM temporel (Time-Aware LSTM). Ce type d'approche a été ensuite utilisé dans (Baytas et al., 2017) pour traiter l'irrégularité temporelle des données médicales en introduisant une porte temporelle supplémentaire sur le réseau LSTM afin que le processus de mise à jour soit au courant de ces intervalles. Il a ensuite été adapté par (Men et al., 2021) afin d'effectuer une prédiction multi-maladies avec un LSTM temporel bidirectionnel.

3 Modèle proposé

Afin d'effectuer des prédictions multi-label sur les actes de soins médicaux, tout en tenant compte de l'hétérogénéité présente dans nos données, nous proposons dans la suite un modèle combinant à la fois l'utilisation d'un LSTM temporel ainsi qu'un clustering flou, comme illustré dans la "Fig. 1". La première étape consiste à adopter un clustering flou du portefeuille des assurés afin de récupérer des probabilités d'appartenance aux différents clusters par assuré. Ces probabilités serviront de métadonnées pour notre modèle séquentiel basé sur un 3-stacked Bi-LSTM temporel. Ce dernier prend en entrée les séquences temporelles d'actes de soins consommés. Nous détaillerons dans la suite chaque étape de notre modèle.

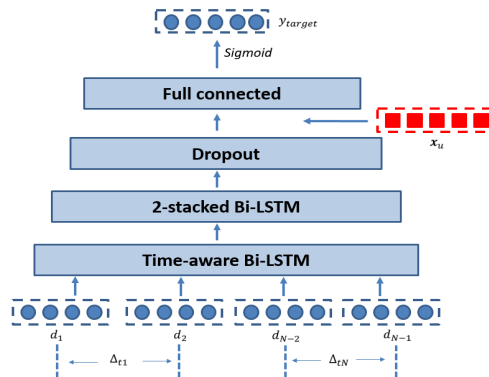


FIG. 1 – Modèle LSTM proposé.

3.1 Clustering flou

Nos données présentent deux niveaux d'hétérogénéité : l'inter-hétérogénéité des différents profils et situations médicales des assurés, et l'intra-hétérogénéité puisqu'un même assuré peut avoir des épisodes médicaux variables dans le temps. Au vu de notre contexte médical, et du

fait que les actes de soins dépendent de la pathologie traitée, nous avons envisagé un clustering qui générerait des clusters pathologiques homogènes regroupant des patients consommant des actes médicaux similaires. Puisqu'un patient peut avoir différents problèmes médicaux qui se chevauchent, une affectation floue aux clusters semble plus précise, de sorte que chaque patient a des probabilités d'appartenance à tous les clusters définis.

Pour ce faire, nous avons utilisé les cartes auto-adaptatives (*Self Organizing Maps - SOM*). Cette technique, proposée par (Kohonen, 2013), est un outil puissant de visualisation et d'analyse de données de grande dimension, qui a été largement utilisé dans divers domaines d'application. La carte auto-adaptative est considérée comme un ensemble de neurones disposés dans une structure de faible dimension de façon qu'il existe des relations de voisinage entre les neurones (Kohonen, 2013). Une SOM fait généralement l'objet d'une étape de clustering dans une approche à deux niveaux, qui vise à regrouper les données en un nombre optimal de classes. Nous commençons donc par réduire la dimension de nos données et les normaliser avant de les projeter sur une SOM, que nous clusterisons à l'aide d'un algorithme Clustering Agglomératif Hiérarchique. Le silhouette score (Rousseeuw, 1987) a été utilisé pour choisir le nombre optimal de clusters. Il en résulte k clusters qui regroupent des patients ayant les mêmes patterns de soins médicaux, ce qui signifie qu'ils sont traités pour des problèmes de santé similaires. Une fois les clusters validés, la dernière étape consiste à appliquer des classifieurs Random Forest sur les k clusters obtenus pour avoir leurs probabilités d'appartenance. Ces probabilités, ajoutées à l'âge et au sexe des patients, serviront de métadonnées à notre modèle LSTM temporel pour capturer les aspects intra et inter de l'hétérogénéité des données et renforcer les prédictions de notre modèle.

3.2 Données et couche d'entrée

Nous implémentons notre modèle sur un jeu de données de 16 Millions d'actes médicaux remboursés, consommés par 150000 assurés entre janvier 2017 et juin 2019. Chaque remboursement est décrit par l'identifiant anonymisé de l'assuré, son âge et sexe, et tous les détails concernant le traitement remboursé comme les dates de début et de fin, la quantité, le prix etc. Les actes de soins sont définis par D labels, représentant la majorité des spécialités médicales telles que la Radiographie, Gynécologie, Chirurgie, Kinésithérapie etc.

Nous regroupons les actes médicaux de chaque patient en veillant à conserver l'aspect temporel, car plusieurs actes médicaux peuvent être consommés dans la même journée, ou s'écarter de plusieurs jours ou mois. Cette préparation est également valable pour la forme tri-dimensionnelle requise pour la couche d'entrée de notre modèle LSTM : Nombre d'échantillons * Taille de séquences * Nombre de variables. Pour la taille des séquences, nous analysons les longueurs des dossiers médicaux des différents patients et expérimentons avec différentes valeurs pour étudier la robustesse des résultats. Une fois la longueur optimale trouvée, nous tronquons les longues séquences et ajoutons du padding au reste. Ensuite, puisque notre ensemble de targets a une taille D d'actes médicaux, un vecteur avec one-hot encoding indiquant laquelle des D spécialités médicales a été consommée représentera chaque séquence. Enfin, nous disposons des trajectoires médicales des 150000 patients de notre portefeuille, dont nous extrayons le dernier mois comme période à prédire et le reste servira d'échantillon d'apprentissage.

3.3 Prédiction multi-label avec 3-stacked Bi-LSTM temporel

Les réseaux de neurones LSTM sont les plus appropriées pour capter l'interdépendance entre les différents actes médicaux composant un dossier médical. De plus, le LSTM bidirectionnel (Bi-LSTM) est une extension des LSTM traditionnels qui entraîne deux LSTM distincts, le premier sur la séquence d'entrée telle quelle et le second sur une copie inversée chronologiquement. Cela fournit un contexte supplémentaire au réseau, ce qui se traduit par un apprentissage plus rapide et plus complet du problème et améliore les performances du modèle sur les problèmes de classification séquence à séquence. Enfin, la recherche (Cui et al., 2020) a prouvé qu'un réseau prédictif plus profond permet une complexité hiérarchique plus importante et serait mieux adapté pour modéliser des données plus complexes. Étant donné que les dossiers médicaux peuvent contenir plusieurs patterns différents, même concernant la même pathologie, cela nous amène à adopter une architecture stacked LSTM.

Dans le domaine de la santé, les antécédents médicaux d'un patient sont une série chronologique séquentielle temporelle avec des corrélations sous-jacentes entre les différents éléments de la trajectoire médicale. Comme mentionné dans l'état d'art, nous avons choisi la variante temporelle du modèle LSTM (Men et al., 2021) comme base pour notre modèle prédictif, car il dispose d'une porte temporelle supplémentaire qui capture l'influence des différents intervalles de temps et contrôle en conséquence le processus de mise à jour.

Notre modèle proposé consiste en un 3-stacked Bi-LSTM temporel : cela signifie une première couche d'un Bi-LSTM temporel dont les états cachés servent d'entrée pour deux couches Bi-LSTM classiques. Enfin, la troisième et dernière sortie de Bi-LSTM constituera la représentation médicale finale, sur laquelle nous appliquons une couche Dropout avant de la concaténer avec les métadonnées préparées (âge, sexe et probabilités de regroupement) x_u . Ces métadonnées nous permettent de capturer à la fois l'hétérogénéité intra et inter des données, représentée par les probabilités de clustering. Nous introduisons ensuite une couche entièrement connectée avec une fonction d'activation sigmoïde pour projeter la représentation combinée sur une distribution de probabilité sur un ensemble d'actes candidats de soins médicaux.

4 Analyse et résultats

4.1 Protocole expérimental

Nous disposons des trajectoires médicales de 150000 de patients, soit 16 millions d'actes médicaux entre janvier 2017 et juin 2019. Lors du déploiement, notre modèle vise à prédire les traitements médicaux dont un patient aura besoin au cours du mois suivant. En conséquence, nous utilisons le mois de juin 2019 pour déterminer les cibles de nos prédictions, et tous les traitements avant cela pour extraire les trajectoires médicales d'entrée. Ensuite, nous divisons nos données en trois ensembles : 70% pour l'ensemble d'apprentissage du modèle, 10% pour l'ensemble de validation et les 20% restants pour l'ensemble de test. L'ensemble de validation est utilisé pour déterminer les hyper-paramètres du modèle à l'aide d'un processus grid search, pour lequel nous essayons différentes combinaisons des valeurs de paramètres suivantes : longueur de la séquence : [25, 50, 100, 150], taille de l'état caché : [32, 64, 128, 256], taux dropout : [0, 1, 0, 2, 0, 3] et taux d'apprentissage : [$2e^{-4}$, $1e^{-3}$, $5e^{-3}$].

La mise en oeuvre de l'expérience a été réalisée sur une carte GPU NVIDIA P100 avec une taille de batch de 128 pour 50 epochs. La taille finale de l'état caché après validation des hy-

perparamètres est de 128 pour les deux premières couches Bi-LSTM et de 64 pour la dernière couche afin de réduire la dimension de l'état caché final, car il sera ensuite concaténé avec les métadonnées. Pour l'évaluation du modèle, les performances de l'algorithme ont été analysées selon trois mesures standard d'apprentissage automatique à partir des données de test : rappel, F1-score et One-error. One-error évalue combien de fois le label le mieux classé n'était pas dans l'ensemble de labels possibles ; sa performance est meilleure plus sa valeur est petite. Le F1-score est une moyenne harmonique entre la précision (pourcentage de labels prédits pertinents) et le rappel (pourcentage de labels pertinents prédits), ayant 1 pour meilleure valeur, et 0 comme pire score. Afin d'obtenir des statistiques fiables sur les mesures de performance, les résultats sont obtenus en moyennant sur 10 itérations des expériences. Pour déterminer si les résultats sont statistiquement significatifs, des t-tests appariés ont été effectués à un niveau de signification de 5%.

4.2 Principaux résultats

Dans cette section, nous rapportons les résultats des performances de prédiction des modèles sur les données de test. Pour établir l'efficacité de notre approche, nous la comparons d'abord à une approche traditionnelle de Binary Relevance (BR) qui transforme le problème de classification multi-label en plusieurs problèmes de classification binaire. Le modèle de base utilisé dans BR était un Random Forest de 100 arbres de décision pour chaque target. Ensuite, puisque la plupart des recherches récentes montrent que les modèles basés sur LSTM surpassent la majorité des méthodes profondes d'apprentissage dans la modélisation de séquences médicales temporelles, nous comparons notre modèle proposé au Bi-LSTM temporel récemment proposé (Men et al., 2021), au 3-stacked Bi-LSTM de base, puis à un 3-stacked Bi-LSTM temporel sans ajouter les métadonnées de clustering flou. Le tableau 1 présente les résultats finaux des performances de prédiction pour chaque modèle.

TAB. 1 – Résultats de performance prédictive. ●/○/= indique que notre approche est significativement meilleure/pire/égale, à un niveau de signification de 5%.

Modèle	<i>F1-score</i>	<i>Rappel</i>	<i>One-error</i>
Binary Relevance – Random Forest	0.582 ●	0.505 ●	0.281 ●
Time-aware Bi-LSTM (Men et al., 2021)	0.618 ●	0.544 ●	0.258 ●
3-Stacked Bi-LSTM	0.615 ●	0.530 ●	0.259 ●
Time-aware 3-Stacked Bi-LSTM sans metadonnées	0.629 ●	0.578 ●	0.252 =
Time-aware 3-Stacked Bi-LSTM avec metadonnées	0.640	0.595	0.247

Trois constats principaux se dégagent de ces résultats. Tout d'abord, nous confirmons que les modèles LSTM surpassent largement l'approche traditionnelle de Binary Relevance, ce qui démontre l'efficacité des techniques d'apprentissage profond dans la modélisation de données médicales temporelles et séquentielles. Deuxièmement, en combinant à la fois les mécanismes du LSTM temporel et des métadonnées de clustering flou, nous voyons clairement l'amélioration significative que notre approche a sur le 3-stacked Bi-LSTM qui surpasse déjà les modèles d'apprentissage profonds existants. En terme de rappel, cette amélioration est aussi importante

que 9% et 6,5% points par rapport aux modèles BR traditionnels et 3-stacked Bi-LSTM, respectivement. Enfin, la comparaison entre notre modèle et les modèles Time-aware Bi-LSTM (Men et al., 2021) et 3-Stacked Time-aware Bi-LSTM sans métadonnées prouve l'avantage d'incorporer les probabilités de clustering flou comme métadonnées. Ce résultat est cohérent avec différentes combinaisons d'ensembles de variables. En bref, nous concluons qu'en capturant l'hétérogénéité intra et inter dans les dossiers médicaux et en l'utilisant pour renforcer un modèle basé sur Bi-LSTM temporel, nous sommes en mesure d'améliorer la prédiction médicale multi-label.

5 Conclusion

Dans cette étude, nous proposons une approche pour effectuer une prédiction multi-label en utilisant des trajectoires d'actes de soins médicaux. Cependant, des défis tels que l'irrégularité temporelle et l'hétérogénéité intra et inter entre les différentes trajectoires médicales des patients compliquent l'utilisation efficace des modèles traditionnels d'apprentissage automatique. Nous résolvons ce problème en développant un modèle d'apprentissage profond 3-stacked Bi-LSTM qui utilise un mécanisme time-aware et le renforce avec des métadonnées de clustering flou. Même si le mécanisme time-aware a été utilisé dans les approches d'apprentissage profond antérieures, le nôtre est parmi les premiers à intégrer l'hétérogénéité des données, par le biais de probabilités d'appartenance à des clusters flous, dans un modèle time-aware pour la classification multi-label dans le domaine de la santé.

Le modèle proposé dans cet article peut être utilisé pour une aide à la décision intelligente dans le monde réel, en particulier en matière d'assurance. Pour démontrer sa faisabilité, nous le validons sur un large portefeuille d'actes médicaux remboursés par une assurance maladie. Les résultats indiquent que notre modèle proposé obtient en effet des performances de prédiction supérieures par rapport aux modèles traditionnels et d'apprentissage profond. Ces prédictions pourraient aider l'assurance maladie à prévenir certains besoins ou traitements médicaux de leurs assurés, voire proposer des garanties plus adaptées à la situation médicale des patients. Une extension significative de ce travail est d'élargir l'entrée de notre modèle par un embedding médical des actes de soins détaillés au lieu des spécialités générales, qui captera plus d'informations sur l'état médical du patient et renforcera les performances prédictives. Une fois l'espace de variables d'entrée étendu, nous aimerions également tester l'efficacité d'une couche d'attention supplémentaire qui permettra d'améliorer les performances prédictives du modèle en capturant l'interdépendance entre les différents éléments d'une séquence médicale, et ainsi pouvoir sélectionner les variables plus pertinentes pour chaque prédiction.

Références

- Baytas, I. M., C. Xiao, X. Zhang, F. Wang, A. K. Jain, et J. Zhou (2017). Patient subtyping via time-aware LSTM networks. In *KDD*, pp. 65–74.
- Choi, E., M. T. Bahadori, A. Schuetz, W. F. Stewart, et J. Sun (2016). Doctor AI : predicting clinical events via recurrent neural networks. In *MLHC*, Volume 56 of *JMLR Workshop and Conference Proceedings*, pp. 301–318.

- Cortez, B., B. Carrera, Y. Kim, et J. Jung (2018). An architecture for emergency event prediction using LSTM recurrent neural networks. *Expert Syst. Appl.* 97, 315–324.
- Cui, Z., R. Ke, Z. Pu, et Y. Wang (2020). Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values. *CoRR abs/2005.11627*.
- Díez-Sanmartín, C. et A. S. Cabezuelo (2020). Application of artificial intelligence techniques to predict survival in kidney transplantation : A review. *Journal of Clinical Medicine* 9.
- Hochreiter, S. et J. Schmidhuber (1997). Long short-term memory. *Neural Comput.* 9(8), 1735–1780.
- Hong, L., M. Luo, R. Wang, P. Lu, W. Lu, et L. Lu (2018). Big data in health care : Applications and challenges. *Data Inf. Manag.* 2(3), 175–197.
- Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks* 37, 52–65.
- Men, L., N. Ilk, X. Tang, et Y. Liu (2021). Multi-disease prediction using LSTM recurrent neural networks. *Expert Syst. Appl.* 177, 114905.
- Nam, J., E. L. Mencía, H. J. Kim, et J. Fürnkranz (2017). Maximizing subset accuracy with recurrent neural networks in multi-label classification. In *NIPS*, pp. 5413–5423.
- Orphanou, K., A. Stassopoulou, et E. Keravnou (2014). Temporal abstraction and temporal bayesian networks in clinical domains : A survey. *Artif. Intell. Medicine* 60(3), 133–149.
- Rousseeuw, P. J. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53–65.
- Wen, T., M. Gasic, N. Mrksic, P. Su, D. Vandyke, et S. J. Young (2015). Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *EMNLP*, pp. 1711–1721. The Association for Computational Linguistics.
- Zhu, Y., H. Li, Y. Liao, B. Wang, Z. Guan, H. Liu, et D. Cai (2017). What to do next : Modeling user behaviors by time-lstm. In *IJCAI*, pp. 3602–3608.

Summary

Medical prevention is a very important aspect of healthcare informatics research through the prediction of medical events. In this work, we propose a deep learning approach to perform multi-label prediction on acts of medical care. The proposed approach utilizes a time-aware long short-term memory network and extends it with additional information from a fuzzy clustering of the same portfolio. The former mechanism (time-aware) is used to handle the temporal irregularity between the elements of a medical trajectory whereas the latter mechanism (fuzzy clustering) assists in modeling the heterogeneity among patients and treatments. Using a large portfolio of reimbursed medical records (over 16 million consumed acts of medical care) by a healthcare insurance in France, we show that our approach outperforms traditional and deep learning methods in medical multi-label prediction. Our work has implications for supporting medical prevention and more broadly improving the quality of healthcare service and insurance.