*MEDINFO 2023 — The Future Is Accessible*
*J. Bichel-Findlay et al. (Eds.)*

*doi:10.3233/SHTI231240*

# Machine Learning Model to Extract Malnutrition Data from Nursing Notes

Mohammad ALKHALAF[a,b,1], Mengyang YIN[c], Chao DENG[d], Hui-Chen (Rita) CHANG[e], Ping YU[a]

[a] *School of Computing and Information Technology, University of Wollongong, Australia*
[b] *School of Computer Science, Qassim University, Qassim, 51452, Saudi Arabia*
[c] *Opal Healthcare, Level 11/420 George St, Sydney NSW 2000, Australia*
[d] *School of Medical, Indigenous and Health Sciences, University of Wollongong, Australia*
[e] *School of Nursing and Midwifery, Western Sydney University, Penrith NSW 2751 Australia*

**Abstract**. Malnutrition is a severe health problem that is prevalent in older people residing in residential aged care facilities. Recent advancements in machine learning have made it possible to extract key insight from electronic health records. To date, few researchers applied these techniques to classify nursing notes automatically. Therefore, we propose a model based on ClinicalBioBert to identify malnutrition notes. We evaluated our approach with two mainstream approaches. Our approach had the highest F1-score of 0.90.

**Keywords**: Natural language processing, malnutrition, nursing progress notes

## 1. Introduction

Malnutrition is a serious health problem in older people [1]. Thus, it is essential to address the nutrition of older people. In Australia, many residential aged care facilities have introduced electronic health records (EHR) which include free-text nursing notes.

The advancement of natural language processing (NLP) technique has afforded us with the opportunity to extract key information and from nursing notes [2]. Previous work on application of NLP in identifying malnutrition notes [3] has identified two challenges: 1) an inability of the machine learning model to distinguish between planned and unplanned weight loss; and 2) not every unplanned weight loss is significant. In this project, we introduce an NLP model based on ClinicalBioBert [4] to identify malnutrition-related notes. We aim for a model that overcomes the above two challenges.

## 2. Methods

This study was approved by the Human Research Ethics Committee at the University of Wollongong. Dataset was obtained from aged care organisation in Australia. It consists of 4445 de-identified residents' information and 1,616,820 nursing progress notes.

The approach to building the NLP model consists of two steps. *First step* is data pre-processing to develop the training data set. We applied ScispaCy UMLS named entity recognition model to identify notes with variables describing weight loss or

[1] Corresponding author: Ping Yu, UOW, Wollongong, Australia, email: ping@uow.edu.au.

malnutrition. After deep analysis by three domain experts with inter-rater reliability of 90%, each note was labelled as either a malnutrition or non-malnutrition related note. If a note states that a resident was malnourished, had an unintentional loss of more than 5% of body weight or unplanned weight loss of more than 3kg in a month [1,5], then that note was considered a malnutrition related note. This process resulted in labelling 628 malnutrition notes and 5000 non-malnutrition notes. S*econd step* started by pre-processing the dataset, and split it into training, validation and testing sets (70-15-15). We utilised ClinicalBioBert and Bert-base models. One main challenge was limitation of token length. Both models do not allow more than 512 tokens for each input sequence; however, many of the notes in our dataset consist of more than that. We addressed this limitation by dividing long notes into three 512 pieces and then take mean pooling output of that piece and then calculating the average model output for the three pieces.

## 3. Results

**Table 1**. Results of malnutrition notes classification model after testing each model on the test dataset

| Model | Runs | Precision | Recall | F1-score |
|---|---|---|---|---|
| Bert-base | Best | 0.84 | 0.84 | 0.84 |
| | Average results of 3 runs | 0.82 | 0.84 | 0.83 |
| ClinicalBioBert | Best | 0.79 | **0.97** | 0.87 |
| | Average results of 3 runs | 0.79 | 0.87 | 0.83 |
| Our approach (1536 tokens) | Best | **0.89** | 0.91 | **0.90** |
| | Average results of 3 runs | **0.87** | **0.88** | **0.87** |

Hyperparameters: learning rate of 5e-5, Adam optimiser, 8 epochs, batch size of 4, and drop out of 0.15

## 4. Conclusions

Availability of NLP and EHR promises brighter future for health informatics. We successfully designed a model to identify malnutrition notes from nursing notes. In the future, we are planning to build a model to predict malnourishment in advance.

## References

[1] Agarwal E, Ferguson M, Banks M, Bauer J, Capra S, Isenring E. Malnutrition coding shortfalls in A ustralian and N ew Z ealand hospitals. Nutr Diet. 2015 Mar;72(1):69-73, doi: 10.1111/1747-0080.12116.

[2] Zhu Y, Song T, Zhang Z, Deng C, Alkhalaf M, Li W, Yin M, Chang HCR, Yu P. Agitation Prevalence in People With Dementia in Australian Residential Aged Care Facilities: Findings From Machine Learning of Electronic Health Records. J Gerontol Nurs. 2022 Apr;48(4):57-64, doi: 10.3928/00989134-20220309-01.

[3] Chen T, Dredze M, Weiner JP, Hernandez L, Kimura J, Kharrazi H. Extraction of Geriatric Syndromes From Electronic Health Record Clinical Notes: Assessment of Statistical Natural Language Processing Methods. JMIR Med Inform. 2019 Mar;7(1):e13039, doi: 10.2196/13039.

[4] E. Alsentzer, J.R. Murphy, W. Boag, W.-H. Wang*, et al.*, Publicly Available Clinical BERT Embeddings, *arXiv* **arXiv:1904.03323** (2019).

[5] Australian Government, QI Program quick reference guide RACF – Unplanned weight loss, (2022)