Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240726

Fill in the Blank, MIMIC the Procedure: Tracking Patient Record for Medical Data Reconstruction

Sujung LEE[a,b], Daechul SEO[b] and Taehoon KO[a,1]
[a] *Department of medical informatics, College of Medicine, The Catholic University of Korea, Republic of Korea*
[b] *Medical AI LAB, MODULABS, Republic of Korea*
ORCiD ID: Sujung Lee https://orcid.org/0000-0001-6151-2497,
Daechul Seo https://orcid.org/0000-0002-0523-9294,
Taehoon Ko https://orcid.org/0000-0002-4045-0036

Abstract. This study addresses the missing data problem in the large-scale medical dataset MIMIC-IV, especially in situations where intubation-extubation events are paired. We employed a strategy involving patient scenario works that checked the temporal order and logical links of intubation/extubation data, and seven reconstruction rules for handling missing values. Through this, we reduced the overall loss rate from 36.89% (3321 records) to 13.37% (1204 records) and achieved a 37.26% data increase (+2117 records) compared to before reconstruction(6582).

Keywords. Medical data reconstruction, missing value, patient tracking

1. Introduction

Large-scale medical datasets like MIMIC-IV [1] are essential for medical research, but they often suffer from incompleteness due to retrospective data collection. Clinically, extubation requires a prior intubation event to be recorded, but discrepancies between intubation and extubation events existed in the dataset, likely due to factors such as EMR recording errors. This study addresses the issue of data loss in medical datasets, particularly missing values in temporal data. Traditional methods can handle missing numerical values [2], but replacing temporal data requires clinical domain knowledge and logical inference. We propose a strategy to mitigate this loss by designing patient scenarios, establishing domain knowledge-based reconstruction rules, and filling in missing values in tasks such as pairing intubation and extubation times.

2. Method

We propose a strategy to mitigate data loss by reconstructing missing values based on the understanding of the actual clinical process and time pairs. The proposed method consists of the following steps: 1) designing patient scenario works that checked the

[1] Corresponding Author: Taehoon Ko, PhD; E-mail: thko@catholic.ac.kr.

temporal order and logical links of intubation/extubation data, 2) establishing seven domain knowledge-based reconstruction rules, and 3) filling in missing values by applying the rules. Through this process, we constructed a paired dataset of 9903 records and classified each record as Extubation Failure (reintubation within 48 hours), Extubation Success or Death. A visual representation of this strategy is available online.

The seven reconstruction rules are briefly listed as follows: RULE#1: If a discharge is not due to death, it is considered a recovery; RULE #2: If possible, substitute values from the "ventilation" table; RULE#3: If extubation is the last procedure, fill in the extubation value with the discharge time if there's no death, else with the time of death; RULE#4: If death occurs more than 48 hours after extubation, categorize as extubation non-failure; RULE#5: If the death occurs within 48 hours of extubation, or if the time of death is earlier than the extubation time, categorized as death; RULE#6: If the time to reintubation is not calculated due to a missing value, replace it with the value from the ventilation table; RULE#7: If cannot be replaced by the value in the ventilation table, check the time difference between the intubation time or extubation time of the current row and the next row, and if the value exists within 48 hours, classify it as extubation failure. If the value does not exist, it is categorized as "Unclassifiable."

3. Results, Discussion and Conclusions

The strategy map was applied to fill missing values in the paired dataset. As a result, complete records increased by 37.26% (+2117 records) compared to before reconstruction (6582 records). Records classified as Extubation Failure, Extubation Success, and Death increased by 1300, 218, and 563, respectively, while Unclassification category decreased from 2294 to 198.

Although the reconstruction process has significantly mitigated the issue of missing temporal data, the study focused only on intubation-extubation pairing, which makes it difficult to claim the procedure's generality. Moreover, due to the absence of actual data, we must rely solely on the logical validity of the reconstructed data. Future research should aim to broaden the scope of this methodology to further enhance the completeness of medical datasets.

Acknowledgement: This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI21C1074). Additionally, it received backing from the Brian Impact Foundation, a non-profit organization dedicated to the advancement of science and technology for all.

References

[1] Johnson AE, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023 Jan 03;10(1):1. doi:10.1038/s41597-022-01899-x.

[2] Zhang Z. Missing data imputation: focusing on single imputation. Ann Transl Med. 2016 Jan;4(1):9. doi: 10.3978/j.issn.2305-5839.2015.12.38.