# Data Engineering to Support Intelligence for Precision Medicine in Intensive Care

1st Nasim Sadat Mosavi
Information systems and technology
*(Algoritmi research center)*
University of Minho
Guimaraes, Portugal
0000-0002-6153-2524

1nd Manuel Filipe Santos
Information systems and technology
*(Algoritmi research center)*
University of Minho
Guimareas,Portugal
0000-0002-5441-3316

***Abstract*—this paper aimed to present the unique data engineering work for dealing with fragmented and infrequent data collection and to integrate data from Intensive Care (ICU) with other resources. The outcome of this data processing supports the development of an Intelligent Decision Support System for Precision Medicine (IDDS4PM) by providing the possibility to analyze all the clinical events in one platform from the date/time of admission. Thus, to obtain the precise treatment, whole clinical data will be considered regardless of the diversity of data sources and frequency of data creation.**

***Keywords— data processing, data engineering, precision medicine, optimal decision making, intensive care.***

***Introduction***

Although there has been a wide contribution to dealing with limitations associated with clinical data processing, the challenges of infrequent data collection, quality issues, data integration, and dimensionality are still under study and investigation. The fact that clinical data is diverse, fragmented, heterogeneous, and unstructured, and has been generated from various resources has made the data processing phase more difficult for developing data-driven solutions. Particularly these limitations are much highlighted for developing intelligent systems such as Precision Medicine (PM) where this method needs efforts to consider whole individual patient data for optimal clinical decision-making. [1]–[3].
According to the U.S National Library of Medicine, PM is a new approach to disease treatment and prevention that takes into account individual variabilities in genes and factors such as environments and lifestyle [4]–[6].

There is no doubt that PM as a new way of decision-making pioneers the shift from the old protocol of clinical decision-making; one-size-fits-all (same treatment for patients with similar symptoms) to "patient-like-me " [7]. Where this replacement will potentially solve the problems associated with the cost of over-treatment, poor quality of healthcare services, and less effective treatment [8], [9]. However, to adopt a such solution, there have been major limitations in the context of clinical data management (data processing, integration, interoperability, privacy ), and a supply of advanced scientific and medical commitment along with the employment of advanced technologies are essential [10], [11].

This paper aimed to present the result of data engineering on 12 datasets from the ICU and other divisions. The first part discusses the scientific gaps in the context of data management and presents the current solutions. After that, the result of data analysis in terms of data acquisition, quality, variables, and records is addressed. Moreover, in the data processing section (third part), initial data manipulation and the novel data processing work are discussed. Finally, the conclusion presents the summary and future works.

## I. BACKGROUND

### *A. Clinical data processing; gaps and limitations*

Although employing an intelligent clinical decision-making system (PM) needs sustainable cooperation between stakeholders, setting new policies and regulations for data interoperability, privacy, and security, aspects related to the maturity, validity, and reliability of relevant clinical practices are also significant. Best practices and valid projects in scale, indirectly boost problem-solving associated with technical areas such as data processing aspects [12],[1].

One of the key limitations that clinical data processing has faced, is in the context of data acquisition and integration which resulted in poor data synchronization [13]. The lack of availability of an integrated data platform causes an isolated and fragmented clinical analysis. Thus, useful data, trends, and information are not able to incorporate into a single model for further clinical actions [14]. Moreover, the fact that data that comes from various resources has a diverse frequency and different time granularity leads to ambiguous data correlation[13].

One common situation is when various physiological indicators in ICU are continuously and regularly monitored and stored: resulting in huge amounts of data collection. To analyze a patient's clinical status and release the best treatment, this data must be considered in integration with other data generated during the admissions of the patient in other sections. This fragmented and infrequent data generation along with other clinical aspects, outliers, and abnormal data, present bias, and in many cases, related data must be ignored or filtered in modeling and further studies [15].

One example is data generated from bedside monitoring which has high frequency while laboratory results are taken irregularly. Based on that, aspects such as frequency and regulations of data generation have a remarkable impact on the performance of the data processing phase [11]. Furthermore, dealing with data quality is another point that depends on major factors such as the assessment of a patient's condition by the clinical team, misinterpretation of the original document, and mistakes in manual data entry[9]. In addition, Medical Waveforms (MW) such as electrocardiograms and electroencephalograms, which are widely utilized in the physiological examination, caries random noise and gaps that influence the quality of data[16].

979-8-3503-2028-2/22/$31.00 ©2022 IEEE
DOI 10.1109/CSCI58124.2022.00342

Therefore, applying suitable methods to manage missing cells is another important task in data cleaning and preparation [17],

As a result, despite the promising role of advanced technologies (Big Data analytics, Machine learning, data mining), and a large number of peer-review articles, limited applications have been used to overcome those aspects. Future efforts should be done to validate the knowledge extracted from clinical data and implement it in clinical practice[18], [19].

Major studies that contributed to offering a practical framework are fragmented and carry limitations. For instance, the "attention scores" measure feature importance, and this method is complex and applicable to nonlinear. [20]. Another practice in clinical data processing, applied hourly aggregation to tackle the challenges of infrequent data registration in the ICU. In this research work, data integration was limited to time series physiological data and laboratory results [20].

Furthermore, dimensionality and velocity are other aspects that must be taken into account in data processing. The TDA method uses algebraic topology to analyze big data by reducing the dimension. This technique is effective for geometric representations and extracting patterns. Moreover, concerning data velocity, the "anytime algorithm" was introduced. This solution is effective for time series data to learn from data streaming. However, the performance of this method depends on the amount of computation they were able to perform.

In addition, although GNMTF is an effective way to solve the challenges associated with clinical data variety (with heterogeneous data), the number of data types for integration influences complexity [19], [21]

In conclusion, existing works have addressed specific challenges and mostly have employed time series data for validation.

## II. DATA COLLECTION AND ANALYSIS

As table 1 shows, there are 12 datasets. The "vital sign" dataset with 43, 9025 records and 108 variables includes biological indicators from ICU. Furthermore, "Lab Result" with 11,3320 records and 9 variables consists of laboratory results. "Procedure", with 911 records and 6 variables has data associated with an action prescribed by doctors. In addition, "SOAP" with 2435 records and 8 variables, keeps key data about the SOAP framework (Subject, Object, Assessment, Plan). In addition, the gravity score or "saps" with 176 records and 6 variables includes data about the level of gravity. Moreover, "Glasgow" carries 861 records and 6 variables about the consciousness status of each patient. Furthermore, data in the "diagnosis" table with 124 records and 9 variables are about signs and symptoms. Additionally, "prescription of medicine" includes data about medications prescribed by the clinician, "administration of medicine", with 993496 records and 17 variables associated with drug administration. Intervention actions are presented in the "intervention" table and "admin-discharge" includes ICU data on admission and discharge. Finally, a reference dataset has an episode/process number. The episode number is a clinical event number and the process number is patient identity.

TABLE1. DATA COLLECTION

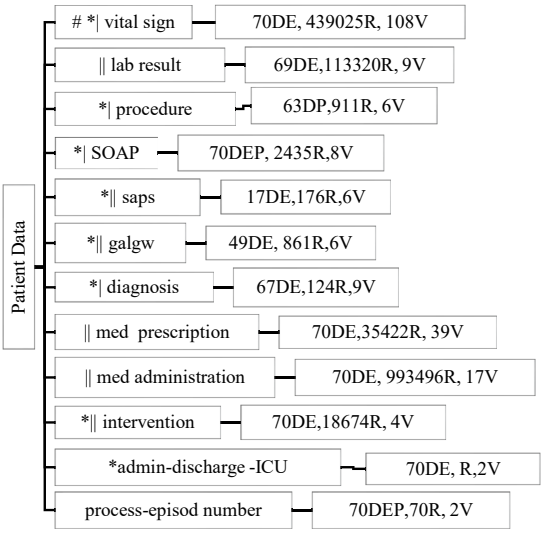



Table 2 describes the symbols used in table 1. According to table 2, only "vital sign" includes time series data. Moreover, tables marked by | ("vital sign", "procedure", "SOAP" and "diagnosis") have time or date of admission, and others with ||, include both (time and date). In addition, * shows that data is associated with ICU (all tables except "med_prescription" and "med administration"). R shows the number of records and V means the number of variables. Moreover, two variables include distinct values whether "Process Number" (DP) or "Episode Number" (DE). According to tables 1 and 2, "procedures" and "SOAP" have DP (distinct process number) and other tables have DE (distinct episode number).

TABLE2. REFERENCE TO TABLE 1

| symbol | description |
|---|---|
| # | Time series data |
| \|\| | Includes time and date |
| \| | Includes time or date |
| * | Intensive care data |
| DE | Distinct Episode number |
| DP | Distinct Process Number |
| R, V | Rows, Feature |

## III. DATA ENGINEERING TO SUPPORT THE ADOPTION OF PM

To solve the limitations associated with data integration, and infrequent data collection and to create an independent platform for developing IDSS 4 PM, in addition to the initial data preprocessing tasks, a novel data engineering task was applied to each dataset.

### *A. Initial data preparation*

According to table 3, initial data preparation consists of major performances such as assessing the quality of data in terms of null values. To deal with missing cells, for some datasets (vital sign, problem) we deleted columns with more than 90% missing values. Moreover, the majority of missing cells were eliminated. In the vital sign dataset, the average value of the previous and next cells was calculated to fill in missing cells. Also, for some datasets such as lab results, we have signed missing cells to decide later in the modeling part.

Moreover, in each dataset feature engineering (extraction, construction, and selection; we extracted time and date from the timestamp and conditional columns were created for datasets such as vital signs, glasgw, and sapsi to identify a patient at risk considering the minimum and maximum boundary of the values associated with biological indicators.

In addition, data collected from biological sensors (vital sign dataset) was aggregated based on hourly circumstances. This task solved the problem associated with infrequent data registration. In the next step, we compared the date of each clinical transaction for every single dataset with the associated date of admission and discharge. In this comparison, we considered episode and process numbers. Finally, each dataset was grouped according to the episode number and sorted based on the date | and time of clinical transactions.

TABLE3.INITIAL DATA PREPARATION| DATA ENGINEERING

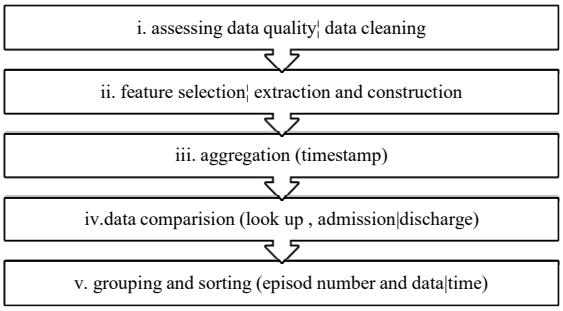
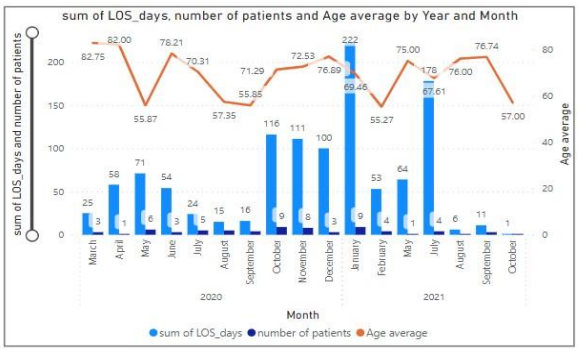



Figure 1, shows an example of an analytical dashboard based on initial data preparation. We used admission and discharge features to create the Length of Stay (LOS) in ICU. In this chart, the total number of patients and LOS are presented. Moreover, the average age level is shown. According to this figure, the maximum LOS was identified in January 2021 (222 days) with 9 patients where the average age level is identified as 69.45. Similarly, in October 2021, the minimum number of the patient was only 1 with a 1-day stay in ICU. The patient was 57 years old.

FIGURE 1. AN EXAMPLE




## B. *A novel method in the clinical data processing*

A new column called Time Slot (TS) is created with a specific logic to transform date and/or time into a unique identity. This value is a pointer to each clinical event for each patient. This type of data engineering is essential for the next phases (discovery, clustering, and prediction). Placing each clinical event associated with each patient provides the possibility to analyze all the clinical transactions from the time/date of admission to discharge regardless of the type of dataset and type of data resources (time series or multivariable). In other words, data will be available with a unique identification in an integrated platform. TS carries several meanings. According to table 4, TS.N.E.Am.Bi. Cj is the representation of this logic where:

TS is Time Slot. Which is the transformation of date| and time into identity.

N is the acronym of the dataset including three letters. For example, “int” is the acronym of the “intervention” dataset. And “vts” is the acronym of the “vital sign” dataset.

E is episode number. Which is unique for each patient.

Am (m=1 ton) is a counter, to check the date and if the date is different from the previous one, make it An+1. For example, the third date of admission makes the counter A3, and the next date of admission would be A4.

Bi (b= 1 ton) is the number of events for the same date. Therefore, on different dates, the counter of B will starts from 1 again. For example, the third event of the first date of admission would be A1.B3.

Finally, Cj shows the number of parallel or sequential events. This number will be changed by checking the time of the clinical event.

TABLE4.CLINICAL DATA ENGINEERING; FORMULA

| TS | N | E | Am | Bi | Cj |
|---|---|---|---|---|---|
| TS | int | 200073 | 1 | 2 | 1 |
| **TS** | Time Slot | | | | |
| **E** | Episode number | | | | |
| **Am** | Counter for the days of clinical events | | | | |
| **Bi** | Counter for the number of clinical events | | | | |
| **Cj** | Counter of the sequence\|parallel events | | | | |
| **m, i,j** | m,i,j = n to n+1 (n !=0) | | | | |
| **Example** | TS.int.200073.1.3.2 | | | | |

Table 5 shows an example of a patient with episode number 20073 from the “intervention” dataset. Where date and time of admission are taken into consideration to create TS for each clinical event.

Based on the proposed data engineering method, in 2020/03/22 there are three different interventions, therefore A is “1” because the first three events happened on the same date. Moreover, B = j+1 (j=1) for every three events and B is an incremental value. Furthermore, since all three events were registered at the same time (20:16), C shows the same number (“1”) for all three events. meaning that there were three parallel interventions. Based on that justification, the first three unique TS are “TS.int.20073.1.1.1”, “TS.int.20073.1.2.1” and “TS.int.20073.1.3.1”. Similarly, there are seven events in 2020-03-23, and for all of them A = 2, because those events are associated with the second day of admissions. Moreover, B= n+ 1 (n0= 1 & n+1<=7). In other words, B has a value from 1 to 7, showing that there are 7 various interventions on the second day of admission. Furthermore, the first three interventions are sequential, meaning that they happened at three different times time1= 03.00, time2=07:00, and time3= 2016. Where the third, fourth, fifth, and sixth are parallel, thus the value of C says

the same (“3”). Finally, the event registered at 23:00 has changed the C value from 3 to 4 and the “TS.int.20073.2.7.4” presents the seventh intervention event on the second day of admission (2020/03/23), where this intervention is the fourth one.

TABLE5. DATA ENGINEERING; EXAMPLE, DATE AND TIME

| date | time | intervention | TS |
|---|---|---|---|
| 2020-03-22 | 20:16 | Avaliar ingestão nutricional | TS.int.20073.1.1.1 |
| 2020-03-22 | 20:16 | Vigiar eliminação urinária | TS.int.20073.1.2.1 |
| 2020-03-22 | 20:16 | Aplicar creme | TS.int.20073.1.3.1 |
| 2020-03-23 | 03:00 | Posicionar | TS.int.20073.2.1.1 |
| 2020-03-23 | 07:00 | Posicionar | TS.int.20073.2.2.2 |
| 2020-03-23 | 20:16 | Vigiar eliminação urinária | TS.int.20073.2.3.3 |
| 2020-03-23 | 20:16 | Aplicar creme | TS.int.20073.2.4.3 |
| 2020-03-23 | 20:16 | Aliviar zona de pressão através de dispositivos | TS.int.20073.2.5.3 |
| 2020-03-23 | 20:16 | Aplicar dispositivos de prevenção de úlcera de pressão | TS.int.20073.2.6.3 |
| 2020-03-23 | 23:00 | Posicionar | TS.int.20073.2.7.4 |

Table 6 shows another example of the applied such data engineering on a diagnosis dataset. In this dataset, we do not have the time of the clinical transaction. Thus the last number in the TS formula is zero. In that table records of two episode numbers (associated with two patients) are presented. The episode number “20012347” has two clinical events on two different dates, so the first TS is “TS.dag.20012347.1.1.0”. where the “dag” is the acronym of diagnostic, after that the episode number is mentioned and the numbers 1.1.0 shows that this clinical event is the first action associated with the first day on 14/04/2020 in ICU and 0 shows that there is no time registered for this transaction. The second record of the same episode number involved a new event on 04/05/2020 (the second day of stay at the ICU) and the TS is “TS.dag.20012347.2.1.0”.

Accordingly, the next episode number has three clinical events at ICU which happened on three different dates. Therefore the first counter after the episode number is specified as 1,2 and 3. Because each day has one event, the second number has remained as 1. Finally, the last number (0) shows that time of the event has not been recorded.

TABLE6. DATA ENGINEERING; EXAMPLE, DATE WITHOUT TIME

| Episod | date | diagnosis | TS |
|---|---|---|---|
| 20012347 | 14/04/2020 | implante ou enxerto vascular NCOP | TS.dag.20012347.1.1.0 |
| 20012347 | 04/05/2020 | aterosclerose de arterias nativas das extremidades, com ulceracao | TS.dag.20012347.2.1.0 |
| 20012892 | 20/04/2020 | Pneumonia bilateral | TS.dag.20012892.1.1.0 |
| 20012892 | 22/04/2020 | Falencia respiratoria | TS.dag.20012892.2.1.0 |
| 20012892 | 04/05/2020 | InsuficiÃªncia respiratÃ³ria aguda grave | TS.dag.20012892.3.1.0 |

## IV. CONCLUSION & FUTURE WORKS

This paper presents the initial result of data engineering on 12 datasets. The majority of data were collected from ICU such as “vital signs”, “procedures”, “SOAP”, “diagnosis”, “saps”, “glasgw”, “intervention”, “admissi_dischcharge” and other data associated with a medical prescription, medical administration and laboratory results were taken into the consideration. Each dataset includes “episode number” which shows the identity of a particular clinical event associated with “process number”. Based on that “process_number” presents a unique identification for each patient.

Moreover, the date/ time of the patient’s admission to various clinical departments is registered in each dataset.

This data engineering work introduced a formula to point to each clinical event with a unique identity by considering the date/time of the clinical event. This performance in addition to initial data processing such as aggregation, feature engineering, data cleaning, and extraction successfully transformed records into independent data from the frequency of data acquisition. The solution has provided synchronizations between various types of data regardless of the type of resources, whether data is time series or not, and has provided an integrated platform for analyzing all clinical data with unique event identification. This ID includes episode number, acronym of the dataset, count of the days of clinical actions, count of the events on the same day, and count of sequence/parallel events.

This solution not only provided a unique time sequence platform for analyzing the whole clinical background from admission to discharge but also, having sync data in a unique platform will facilitate the development of future works for the deployment of an IDSS4PM. Clustering similar patients, predicting the future patient medication / clinical status, and proposing the optimized treatment are identified as future works.


## ACKNOWLEDGMENT

The work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Projects Scope: DSAIPA/DS/0084/2018


## REFERENCES


[1] N. Sadat Mosavi and M. Filipe Santos, “Adoption of Precision Medicine; Limitations and Considerations,” 2021, pp. 13–24.

[2] X. Liu, X. Luo, C. Jiang, and H. Zhao, “Difficulties and challenges in the development of precision medicine,” *Clin. Genet.*, vol. 95, no. 5, pp. 569–574, May 2019.

[3] J. McPadden *et al.*, “Health care and precision medicine research: Analysis of a scalable data science platform,” *J. Med. Internet Res.*, vol. 21, no. 4, pp. 1–11, 2019.

[4] M. Haque, T. Islam, M. Sartelli, A. Abdullah, and S. Dhingra, “Prospects and challenges of precision medicine in lower-and middle-income countries: A brief overview,” *Bangladesh J. Med. Sci.*, vol. 19, no. 1, pp. 32–47, 2020.

[5] N. S. Mosavi and M. F. Santos, “How prescriptive

analytics influences decision making in precision medicine," *Procedia Comput. Sci.*, vol. 177, pp. 528–533, 2020.

[6] N. S. Mosavi and M. F. Santos, "To what extent healthcare analytics influences decision making in precision medicine," *Procedia Comput. Sci.*, vol. 198, no. 2021, pp. 353–359, 2021.

[7] N. S. Mosavi and M. F. Santos, "Internet of things for precision intensive medicine," *Procedia Comput. Sci.*, vol. 201, no. C, pp. 732–737, 2022.

[8] G. Barros, "Herbert A . Simon and the concept of rationality : Boundaries and procedures," vol. 30, no. March 2009, pp. 455–472, 2010.

[9] S. A. Brown, "Patient similarity: Emerging concepts in systems and precision medicine," *Front. Physiol.*, vol. 7, no. NOV, pp. 1–6, 2016.

[10] C. Kennedy and J. Turley, "Time Series Analysis As Input for Predictive Modeling : Predicting Cardiac Arrest in a Pediatric Intensive Care Unit," *Theor. Biol. Med. Model.*, vol. 8, no. 40, pp. 1–25, 201AD.

[11] P. Y. Wu, C. W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, "-Omic and Electronic Health Record Big Data Analytics for Precision Medicine," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 263–273, 2017.

[12] A. Blasimme, M. Fadda, M. Schneider, and E. Vayena, "Data sharing for precision medicine: Policy lessons and future directions," *Health Aff.*, vol. 37, no. 5, pp. 702–709, 2018.

[13] Y. Zhang, C. T. Silvers, and A. G. Randolph, "Real-time evaluation of patient monitoring algorithms for critical care at the bedside," *Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc.*, pp. 2783–2786, 2007.

[14] C. E. Kennedy and J. P. Turley, "Time series analysis as input for clinical predictive modeling: Modeling cardiac arrest in a pediatric ICU," *Theoretical Biology and Medical Modelling*, vol. 8, no. 1. 2011.

[15] A. A. Seyhan and C. Carini, "Are innovation and new technologies in precision medicine paving a new era in patients centric care?," *J. Transl. Med.*, vol. 17, no. 1, pp. 1–28, 2019.

[16] S. Khadanga, K. Aggarwal, S. Joty, and J. Srivastava, "Using Clinical Notes with Time Series Data for ICU Management," Sep. 2019.

[17] F. I. Adiba, S. N. Sharwardy, and M. Z. Rahman, "Multivariate time series prediction of pediatric ICU data using deep learning," *2021 Int. Conf. Innov. Trends Inf. Technol. ICITIIT 2021*, 2021.

[18] G. Carra, J. I. F. Salluh, F. J. da Silva Ramos, and G. Meyfroidt, "Data-driven ICU management: Using Big Data and algorithms to improve outcomes," *J. Crit. Care*, vol. 60, pp. 300–304, 2020.

[19] N. S. Mosavi and M. F. Santos, "Intelligent Decision Support System for Precision Medicine (IDSS 4 PM)," vol. 3, no. Ic3k, pp. 29–36, 2022.

[20] N. Johnson, S. Parbhoo, A. S. Ross, and F. Doshi-Velez, "Learning Predictive and Interpretable Timeseries Summaries from ICU Data," *arXiv.org*. 2021.

[21] V. Gligorijević, N. Malod-Dognin, and N. Pržulj, "Integrative methods for analyzing big data in precision medicine," *Proteomics*, vol. 16, no. 5, pp. 741–758, Mar. 2016.