Subpopulation Process Mining in the ICU for Patients Diagnosed with a type of Pneumonia

Lisan Graumans
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT

COVID-19 has impacted the ICU of hospitals all over the world, this results in a new kind of society where people must try to reduce to spread of the pandemic to ensure that the ICU can treat all patients. This research is a foundation to conduct further research into pneumonia, with the focus on the ICU and COVID-19. The data from MIMIC III[2] is used to discover the care paths of different subpopulations going through the ICU. The subpopulations consist of patients diagnosed with a form of pneumonia, specifically, viral, bacterial and fungal. The three care paths are compared via visual comparison to detect conformance and deviation. The care paths mostly conform and show only a few deviations. To substantiate the conformance and deviation of the subpopulations, the comparison of twelve biomarkers is included.

Keywords

process mining, care paths, pneumonia, ICU

1. INTRODUCTION

The current ongoing pandemic highlights the importance of the organisation of an ICU. COVID-19 has shown the world how an inordinately increased demand for ICU places can disrupt society. To lay a foundation for further research into the impact of COVID-19 on the ICU, this research will use past data to understand the current organisational aspects concerning the patients diagnosed with pneumonia. The care paths of the patients will be represented by three different process models of three different subpopulations. The subpopulations this research focuses on, are viral, bacterial and fungal pneumonia. The models will provide insights on the procedures the patients will be treated with. The comparison of process models will be expanded with a comparison of twelve biomarkers. The biomarkers will show conformance or deviation from a different point of view.

1.1 Research Questions

This research will answer the following research questions:

1. What does the model for care paths for a patient, diagnosed with a type of pneumonia, admitted to the ICU look like?
2. How do the different care paths deviate and or conform?
   (a) What subpopulations can be identified?
   (b) What different care paths can be identified?

The research questions will be answered with the use of process mining. Process mining includes different techniques to extract information from event logs. The extracted information is in the form of process models. The process models will reflect the path of a patient through the ICU, these models will be compared to detect conformance and deviation. To show the paths of the patients from another point of view, a set of twelve biomarkers, obtained from the same database, are also checked to highlight conformance or deviation.

This paper commences with some background information on process mining and the data, followed by a literature review and an account of the tools used to conduct the research. The methodology gives a step by step overview of the set-up, the results consist of the findings, a discussion and a comparison with the literature. The last section will describe the limitations and future work and the conclusion.

2. BACKGROUND

2.1 Process Mining

Process mining is a term for the multitude of techniques used to extract knowledge from an event-based data-set to analyze this knowledge. The event-based data-set must contain a timestamp, an activity and a case ID to be usable.

Process mining is used to discover, monitor and enhance processes. The *discovery* of processes starts with a usable data-set. From the data-set, a model will be extracted that shows the process from start to finish. Often discovery is the starting point for a broader study. The process model that has been discovered can be compared to the event log or to other models. This is known as *conformance checking* and can be used to detect deviation and conformance. Lastly, the *enhancement* of processes establishes a new model based on the data-set and the process model. The new model can, for example, be used to reorganize the present execution of processes or help to reduce bottlenecks. [6]

2.1.1 Tools

ProM[1] was used to execute the process mining. It is the tool that is most often used in process mining in health.[5]

[1] https://www.promtools.org/doku.php

34th Twente Student Conference on IT Jan. 29nd, 2021, Enschede, The Netherlands.

It is available to everyone and has an extensive amount of plug-ins. The plug-ins used in this research are:

- Convert CSV to XES
- Mine with Inductive visual Miner

2.2 Data

The data used to conduct this research, is data from the MIMIC III database[2]. This database contains anonymized health data. The attributes that are used in this research are the procedures and their sequence number, the biomarkers and their values and the patients and their corresponding diagnosis.

2.2.1 ICD codes

The database uses ICD-9 codes to refer to the diagnoses. The ICD-9 codes are out of use and replaced by the newer version of ICD-10 codes.
In this research, the focus will be on patients in the Intensive Care Unit diagnosed with an ICD-9 code starting with 48. This indicates a type of pneumonia. This focus is chosen because COVID-19 is a form of viral pneumonia.

2.2.2 Source

The source of the data is the Medical Information Mart for Intensive Care III database. The relational database contains data from over 40.000 patients in the ICU from 2001 to 2012 at Beth Israel Deaconess Medical Center. The database consists of 40 tables, examples are admissions, diagnoses, transfer, procedures and more table containing information about medical treatment and organizational processes. [2] In this research, the focus is on medical procedures and biomarkers.

2.3 Biomarkers

Biomarkers are measures that are used to indicate a patient's state of health. The biomarkers give context to a patient's care path and they can be the reason for a certain procedure. This research makes a selection of biomarkers to compare the values of the three subpopulations.

3. RELATED WORK

Related work to this research that highlights the importance of process mining in health is *Process Mining in Healthcare*, in this paper not only the relevance of process mining within healthcare is discussed, but it also makes a distinction between the different kinds of purposes or process mining. The purpose that is most relevant in this research is the exploration of different processes and models to scope the analysis. [3]
Automatic Process Comparison for Subpopulations: Application in Cancer Care is related work that proposes measures to compare different subpopulations.[4] This research uses visual comparison, meaning that the models will be compared with the unaided eye.
Process Mining in healthcare: A literature review[5] provides multiple criteria to distinct this research. Process mining in healthcare can focus on different processes, a processes can be categorized as an organizational process or a medical treatment process. Examples are respectively; assigning a certain number of employees to a certain task and the diagnosis of a patient. Another criterion is the perspective of the process mining[6], within health, there are four perspectives identified, namely, control flow, performance, conformance and organizational.

- Control flow entails the discovery of the actual sequence of processes.
- Performance focuses on points of improvements, for example, bottlenecks or idle time.
- Conformance checks how processes deviate from the pre-established model.
- Organization is used to inspect the combination and use of resources.

This research starts with the control flow of the ICU patients and their medical treatment processes, this will be followed by a conformance check for the three subpopulations.
The control flow of ICU patients has previously proven to be difficult to standardize. Patients admitted to the ICU are severely ill and can have multiple diseases. This results in numerous amount of processes, which are not always related. [1]
This research's focus is on patients admitted to the ICU, the process models will show a low number of procedures compared to the total number of procedures in the database. Otherwise, it is not possible to draw a conclusion based on the process models.

4. METHODOLOGY

4.1 Gaining access to data

The database MIMIC III was selected because of its size, this provides a degree of certainty that if adjusting the scope, there is still enough data to conduct research with. To gain access to the database set up by MIT[2], the user will have to complete nine modules on the ethics of clinical research by MIT, send them information about your research and the contact information of your supervisor. They will approve or decline the request because MIT wants to ensure they do not grant access to those with bad intentions.

4.2 Selecting subpopulations

The three subpopulations are viral, bacterial and fungal pneumonia. Each subpopulation consists of patients diagnosed with the corresponding type of pneumonia. These three subpopulations were chosen in consultation with a medical expert of MST. Three types of pneumonia were decided upon because COVID-19 is also a type of pneumonia, namely viral pneumonia. COVID-19 has caused a great demand for beds in the ICU, due to this, the care paths of patients diagnosed with a type of pneumonia will be the target of this research.

Viral	4800, 4801, 4802, 4803, 4808, 4809,
	4841, 4843, 4845, 4870, 4871, 48801
	48802, 48811, 48812, 48881, 48882
Bacterial	481, 4820, 4821, 4822, 48230
	48231, 48232, 48239, 48240, 48241
	48242, 48249, 48281, 48282, 48283
	48284, 48289, 4829, 4830, 4831, 4847
Fungal	4846

Table 1. The ICD9-codes per subpopulation

4.3 Creating models

The models are created with ProM. Every subpopulation has a process model representing the path through the ICU. All relevant admissions have been used to generate the model. The relevant admissions are those concerning

[2] https://research.mit.edu/integrity-and-compliance/responsible-conduct-research

a patient diagnosed with a type of pneumonia. The activities in the model are the procedures, the timestamp is the sequence of the procedures and the case ID is the admission ID. The models are inductive process models. The inductive process miner repeatedly locates a split in the data-set, detects the correct operator and continues with the remaining data.

4.4 Comparing models

The models are compared by detecting visual conformance and deviation. The visual comparison is done by the unaided eye. The three models, one for each subpopulation, will be compared to each other.

4.5 Selecting biomarkers

The biomarkers have been selected based on a consult with the expert. We assume that these biomarkers will have a relationship with COVID-19. These biomarkers indicate the state of health of a patient. Table 2 shows the biomarkers, together with their unit, that are used in this research. The last one, length of stay, is not a biomarker. The length of stay refers to the duration the patient has been in the ICU.

Biomarker	**Unit**
Alveolar-arterial Gradient	mm Hg
Atypical Lymphocytes	%
Calculated Total CO2	mEq/L
C-Reactive Protein	mg/L
Eosinophils	%
Lactate	mmol/L
Lymphocytes	%
Monocytes	%
Neutrophils	%
pCO2	mm Hg
pO2	mm Hg
White Blood Cells	K/uL
Length of stay	days

Table 2. Biomarkers with corresponding units

4.6 Comparing biomarkers

To compare the biomarkers, the total number of measures have been used to calculate the average and standard deviation per subpopulation. To explain some of the difference, a count of the number of times a biomarker has been measured is added.

The comparison of the average and the standard deviation will introduce a point of view next to the process models.

5. RESULTS

5.1 Creating models

The process models representing the care path of the patients going through the ICU are added to the appendices. Appendix A contains the model of the patients diagnosed with viral pneumonia, appendix B contains the model of the patients diagnosed with bacterial pneumonia and appendix C contains the model of the patients diagnosed with fungal bacteria. The parameters of the process models are the percentages of the total activities included, the percentages indicating the number of paths and the fitness of the models.

5.2 Comparing models

The models are compared with visual comparison. The models show the different care paths of patients in the form of procedures.

	Activities	Paths	Fitness
Viral	0.077	0.8	0.848
Bacterial	0.012	0.8	0.721
Fungal	0.1	0.8	0.717

Table 3. The parameters of the process models

As can be seen in the process models and in table 4, most of the procedures conform. Especially bacterial, this process model does not have a unique procedure. One remark on this process model, the procedure *Cont inv mec ven 96+ hrs* goes unprecedented, this can also still be true for patients with a form of viral or fungal pneumonia. The viral process model has two unique procedures and is most similar to the bacterial process model. The fungal process model has the most unique procedures. Except for the unique procedures, which are deviations, the models mostly conform. Noticeable is the start of care path is in all three cases often a *Temporary tracheostomy.*

	Viral	Bacterial	Fungal
Temporary tracheostomy	x	x	x
Percu endosc gastrostomy	x		
Venous cath NEC	x	x	x
Cont inv mec ven <96 hrs	x	x	
Packed cell transfusion	x	x	x
Hemodialysis	x		
Entral infus nutrit sub	x	x	x
Closed bronchial biopsy	x	x	x
Parent infus nutrit sub	x	x	x
Cont inv mec ven 96+ hrs	x	x	x
Arterial catheterization	x	x	x
Insert endotracheal tube	x	x	x
Spinal tap	x		x
Platelet transfusion			x
Thoracentesis			x
Ven cath renal dialysis			x

Table 4. The table shows the presence of procedures in the process models

5.3 Comparing biomarkers

The total measures for each biomarker have been used to calculate the average and the standard deviation.

Table 5 shows the count of each biomarker per subpopulation, this gives context to the values and explains some of the extremes.

Table 6 shows the average of each biomarker. Each subpopulation has a biomarker that noticeably differs from the other subpopulations. Viral has *Alveolar-arterial Gradient* and *pO2*, which both differ from the values of bacterial and fungal. Bacterial is often the middle-way between viral and fungal, except for the *White Blood Cells.* Fungal deviates the most compared to viral and bacterial. *Atypical Lymphocytes*, *Lymphocytes*, *Monocytes* and *Neutrophils* all show a considerable difference. Table 7 shows the standard deviation of each biomarker per subpopulation. The biomarker averages that stood out in the viral subpopulation do not show a substantial difference. In the case of standard deviation, the *Eosinophils* and *White Blood Cells* deviate. The biomarkers for the bacterial subpopulation do not show any unusual deviation. The fungal subpopulation shows deviations in the same biomarkers noted in table 6.

Biomarker	Viral	Bacterial	Fungal
Alveolar-arterial Gradient	426	3889	109
Atypical Lymphocytes	649	5601	656
Calculated Total CO2	4123	84029	2093
C-Reactive Protein	5	11	x
Eosinophils	733	6478	682
Lactate	132	2854	96
Lymphocytes	759	7192	732
Monocytes	750	6816	705
Neutrophils	720	6179	665
pCO2	4124	84030	2091
pO2	4125	84030	2091
White Blood Cells	352	6486	225
Length of stay	240	3072	66

Table 5. The count of all measures per biomarker and subpopulation.

Biomarker	Viral	Bacterial	Fungal
Alveolar-arterial Gradient	520.16	497.78	499.86
Atypical Lymphocytes	0.78	0.69	1.19
Calculated Total CO2	27.84	27.56	26.13
C-Reactive Protein	48.8	135.18	x
Eosinophils	1.85	1.64	1.24
Lactate	1.96	2.19	2.35
Lymphocytes	16.92	15.65	32.16
Monocytes	6.12	6.82	8.99
Neutrophils	65.61	69.16	45.92
pCO2	46.71	44.82	43.70
pO2	104.13	113.93	111.26
White Blood Cells	10.47	12.31	10.69
Length of stay	8.95	11.66	15.16

Table 6. Average of biomarkers per subpopulation

5.4 Discussion

The table of the count of the measures of the biomarkers was added to indicate the significance of the results. The difference in C-Reactive Protein is not significant because there is almost no data, the same applies to the length of stay.
The bacterial subpopulation includes most ICD9-codes and the most patients, this can be a factor in the conformance checking because viral and fungal have a lot fewer data.

5.5 Comparison to literature

The literature[1] explained the difficulty of creating process models based on the care paths through the ICU. This can also be seen in the parameters used to create the models. A big number of procedures is excluded because the procedure would only happen once to one patient.

6. LIMITATIONS & FUTURE WORK

A limitation of this research lays within the database. The procedures are not timestamped only sequenced. A timestamp indicates the exact time, a sequence only tells the order of execution. The biomarkers are timestamped. It is impossible to determine the direct consequence of a biomarker because it is not possible to determine the exact moment of the biomarker compared to the moment of the procedure.
Another limitation, perhaps also an opportunity, is the ICU itself. As a result of the standardization of a care path through the ICU, a lot of data is lost. The activities that only happen a few times are not visible in the model but they do have an impact on the biomarkers. This can also be presented as an opportunity. The lack of structure within the ICU can make the small amount of, can make a small amount of structure discovered through process mining more valuable.

Biomarker	Viral	Bacterial	Fungal
Alveolar-arterial Gradient	95.36	104.94	101.34
Atypical Lymphocytes	2.66	2.03	3.03
Calculated Total CO2	6.91	6.39	6.43
C-Reactive Protein	38.21	112.84	x
Eosinophils	7.33	4.14	2.93
Lactate	1.41	1.95	2.40
Lymphocytes	20.94	20.62	32.80
Monocytes	6.57	8.73	12.86
Neutrophils	27.51	23.72	32.97
pCO2	13.07	12.43	11.26
pO2	54.80	60.60	50.29
White Blood Cells	16.79	7.37	9.95

Table 7. Standard deviation of biomarkers per subpopulation

Future work can be done in this field to determine the impact of COVID-19 on the organisation of the ICU. the care paths of COVID-19 patients can be compared to those of older pneumonia patients, the latter as described in the research.
subpopulations can also be community-acquired pneumonia versus hospital-acquired pneumonia. Since COVID-19 is mostly community-acquired it will have a high chance of showing conformance with the care paths of patients diagnosed with community-acquired viral pneumonia. This research can create a timeline of the development of different types of pneumonia.

7. CONCLUSION

The research questions, representing the goals of this research, will be answered to conclude this final section.
The first goal of the research was to determine what a care path for a patient, diagnosed with a form of pneumonia, going through the ICU looked like. This question has been narrowed down to a patient diagnosed with a form of pneumonia. The care path can be seen in the appendices, notable about the care path is the low number of procedures included because the patients admitted to the ICU often need more procedures and are diagnosed multiple times with different diseases.
The second goal was to identify different subpopulations and different care paths and to detect conformance and deviation. The bacterial process model shows the most conformance as it has no unique processes. The biomarker that shows deviation is *White Blood Cells.* The viral process model has more deviations than bacterial but still is mostly conform. The biomarkers that show deviation are *Alveolar-arterial Gradient* and *pO2.* Fungal is also mostly conforming, but shows the most deviation with the biggest number of unique procedures. The biomarkers that show deviation are *Atypical Lymphocytes*, *Lymphocytes*, *Monocytes* and *Neutrophils.*
The subpopulations used in this research, viral, bacterial and fungal, especially show considerable different values

of biomarkers. The process models are more conform. To conduct further research into the care paths the subpopulations should be more specific, fewer ICD9-codes per subpopulation or an extra division between hospital-acquired and community-acquired pneumonia.

8. REFERENCES

[1] S. Gupta. Workflow and process mining in healthcare. pages 79–80, June 2007.

[2] S. L. L. L. F. M. G. M. M. B. S. P. C. L. Johnson AEW, Pollard TJ and M. RG. Mimic-iii, a freely accessible critical care database. 2016.

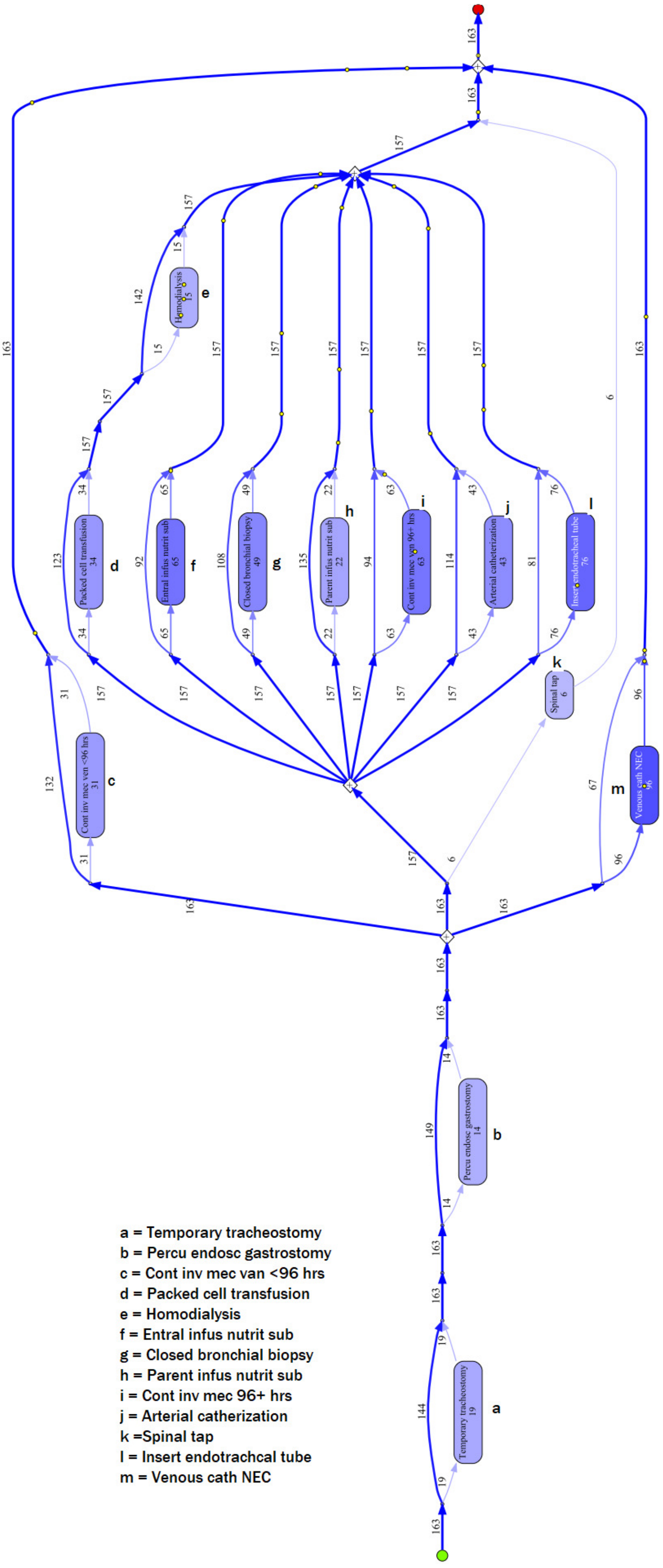
[3] A. v. d. W. M. P. . V. R. J. B. Mans, R. S. Process mining in healthcare : opportunities beyond the ordinary. *BPM REPORTS*, 1326:52, January 2013.

[4] F. Marazza, F. A. Bukhsh, J. Geerdink, O. Vijlbrief, S. Pathak, M. V. Keulen, and C. Seifert. Automatic process comparison for subpopulations: Application in cancer care. *International Journal of Environmental Research and Public Health*, 17(16):5707, Augustus 2020.

[5] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro. Process mining in healthcare: A literature review. *Journal of Biomedical Informatics*, 61:224–236, June 2016.

[6] W. van der Aalst. Process mining. *Communications of the ACM*, 55:76–83, August 2012.

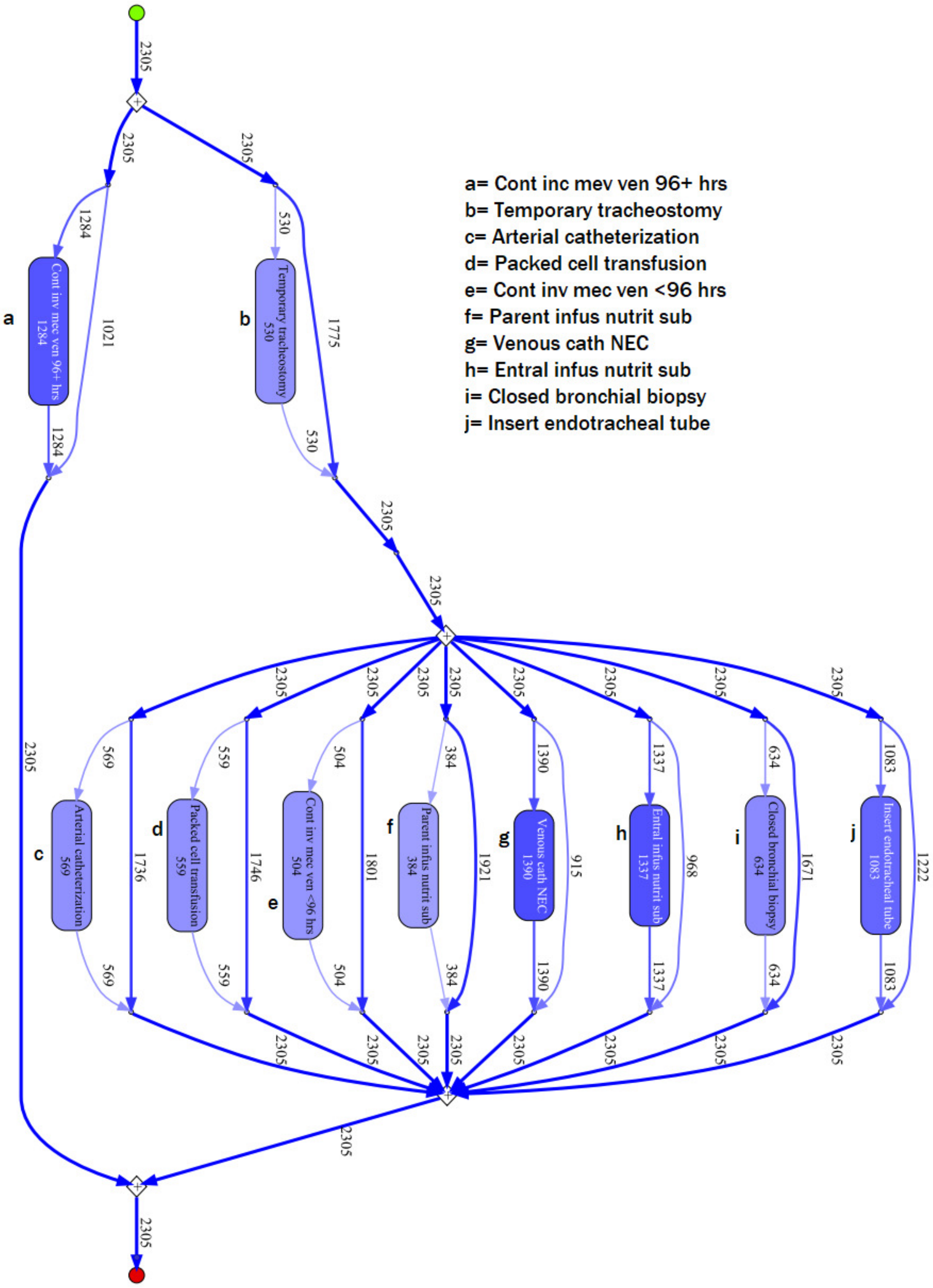
Appendix A. Process model of patients diagnosed with viral pneumonia



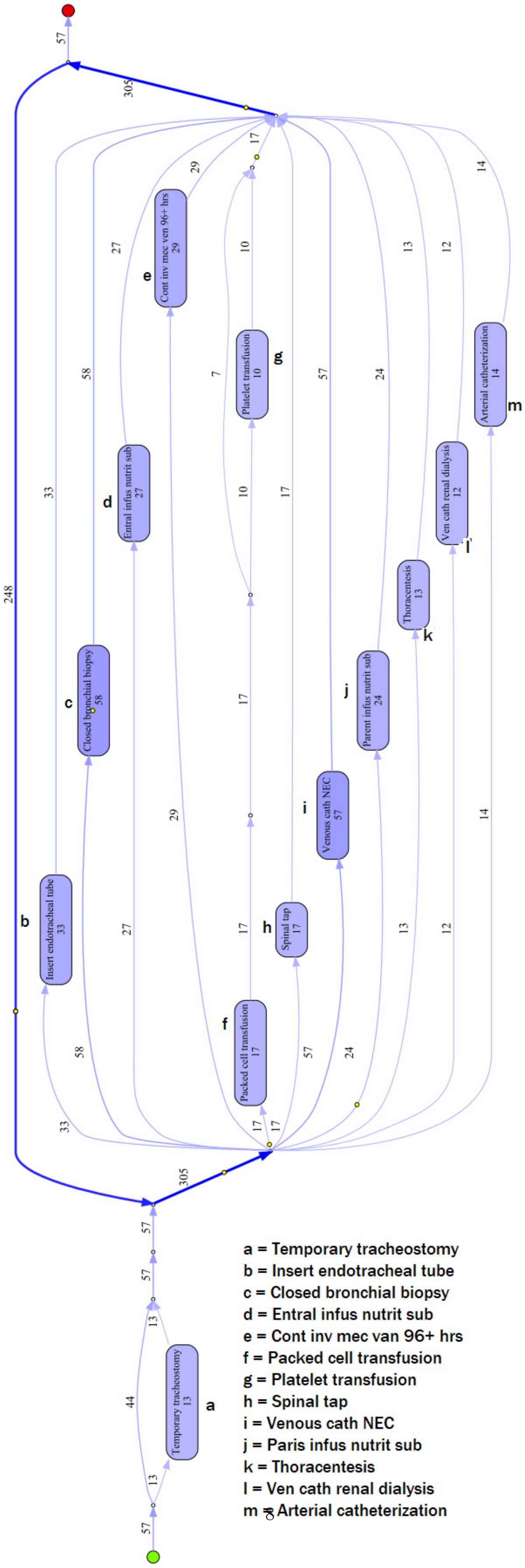
Temporary tracheostomy 19
a
Percu endosc gastrostomy 14
b
Cont inv mec ven <96 hrs 31
c
Packed cell transfusion 34
d
Hemodialysis 15
e
Entral infus nutrit sub 65
f
Closed bronchial biopsy 49
g
Parent infus nutrit sub 22
h
Cont inv mec ven 96+ hrs 63
i
Arterial catheterization 43
j
Spinal tap 6
k
Insert endotracheal tube 76
l
Venous cath NEC 96
m
163
144
149
157
132
142
123
92
108
135
94
114
81
67
6

a = Temporary tracheostomy
b = Percu endosc gastrostomy
c = Cont inv mec van <96 hrs
d = Packed cell transfusion
e = Homodialysis
f = Entral infus nutrit sub
g = Closed bronchial biopsy
h = Parent infus nutrit sub
i = Cont inv mec 96+ hrs
j = Arterial catherization
k =Spinal tap
l = Insert endotrachcal tube
m = Venous cath NEC

Appendix B. Process model of patients diagnosed with bacterial pneumonia



Appendix C. Process model of patients diagnosed with fungal pneumonia


Temporary tracheostomy 13
a
Insert endotracheal tube 33
b
Closed bronchial biopsy 58
c
Entral infus nutrit sub 27
d
Cont inv mec ven 96+ hrs 29
e
Packed cell transfusion 17
f
Platelet transfusion 10
g
Spinal tap 17
h
Venous cath NEC 57
i
Parent infus nutrit sub 24
j
Thoracentesis 13
k
Ven cath renal dialysis 12
l
Arterial catheterization 14
m
57
13
44
13
57
57
305
248
305
57
33
58
27
29
17
17
57
24
13
12
14
10
10
7
17
17
17
29
27
33
58
57
24
13
12
14

a = Temporary tracheostomy
b = Insert endotracheal tube
c = Closed bronchial biopsy
d = Entral infus nutrit sub
e = Cont inv mec van 96+ hrs
f = Packed cell transfusion
g = Platelet transfusion
h = Spinal tap
i = Venous cath NEC
j = Paris infus nutrit sub
k = Thoracentesis
l = Ven cath renal dialysis
m = Arterial catheterization