

# Development of BDMS Utilization Review Technology (BURT): An Artificial Intelligence Tool Using Thai Natural Language Processing to Assess Appropriateness of Hospitalization

Jinhatha Panyasorn, DDS, MSc, CHCQM<sup>1</sup>; Piemchok Banomyong, MD<sup>1</sup>; Kusuma Phetchunsakul, RN<sup>1</sup>; Noppadol Phengpinit, RN<sup>1</sup>; Varut Wiseschinda<sup>2</sup>; Chaiyos Kunanusont, MD, PhD<sup>2</sup>



Jinhatha Panyasorn  
DDS, MSc, CHCQM

## Abstract

**OBJECTIVES:** To develop an effective artificial intelligence (AI) driven platform to optimize the process of assessing appropriateness of hospitalization.

**MATERIALS AND METHODS:** Anonymized data of 22,020 insured-patient admissions in a BDMS network hospital were included to build a prediction model based on a comprehensive guideline for appropriate hospitalization. To develop Thai Natural Language Processing (NLP) model, 77,707 sentences from medical records were used and separated into two datasets, 80% for training and 20% for testing. A combined NLP and rule-based algorithms formed an AI engine and outputs were displayed using a web-based application. An expert panel of five Utilization Management (UM) physicians had several collaborative discussions to fine tune the NLP model, application of clinical criteria, and classification engine. Eventually, NLP model in the latest version (BURT1.1), had satisfactory features with overall higher than 99% accuracy, precision, recall, and F1.

**RESULTS:** Performance of BURT1.1 was assessed using 300 cases randomly selected from the main dataset, against other methods, including concurrent review by UM nurses at the participating hospital, and UM nurses at Bangkok Hospital Headquarters (BHQ). Agreement upon UM Physician Panel consensus was set as one of the performance indicators, and BURT1.1 showed a favorable outcome with the highest rate of agreement (86%) among all the methods. The precision rate was 99% as compared to insurance claim approval status. Additionally, dramatic time savings were achieved with 0.59 second of processing time as compared to 10-15 minutes per case by conventional manual review.

**CONCLUSION:** BURT1.1 should be effectively implemented as an automatic daily tool to screen inappropriate hospitalization. It can immediately identify patients at high risk of inappropriate hospitalization that require further assessment by UM nurse, thus providing feedback to attending physicians on the completeness and quality of documentation, with parallel notification to UM physicians. Ultimately, BURT1.1 can contribute to increase UM efficiency, speeding up the claim process, reducing health care costs due to unnecessary hospitalization, and reduction of claim denials.

**Keywords:** BDMS utilization review technology (BURT), natural language processing (NLP), rule-based algorithm, web application, UM physician, admission criteria, appropriate hospitalization

Utilization management (UM) has been effectively used as one of the approaches to reduce consumption of unnecessary healthcare services and thus helps contain cost. It is particularly important in the health insurance industry, due to moral hazard effect that changes patient and physician behaviors and results in overutilization of resources.<sup>1,2</sup>

A range of utilization management procedures have been deployed by third-party-payers to prevent inappropriate admission. Pre-authorization is a widely-used technique to certify the need for hospitalization and medical care, but it contributes to administrative burden and may delay necessary services. As a result, most insurance companies in Thailand implement a pre-authorization process for non-urgent surgical procedures and high-cost

<sup>1</sup> Corporate Utilization Management and Insurance Services, Bangkok Dusit Medical Services, Bangkok, Thailand

<sup>2</sup> Bangkok Health Research Center, Bangkok Dusit Medical Services, Bangkok, Thailand

\* Address Correspondence to author:  
Jinhatha Panyasorn  
Corporate Utilization Management and Insurance Services  
Bangkok Dusit Medical Services  
2 Soi Soonvijai 7, NewPetchburi Rd,  
Bangkok 10310, Thailand.  
email: Jinhatha.Pa@bdms.co.th

diagnostic procedures only, leaving non-surgical cases to be submitted and adjudicated after the services have already been delivered. Unexpectedly rejected claims can sometimes cause confusion and frustration to patients. Due to the aforementioned reasons, concurrent review is currently a main monitoring measure performed by both insurance companies and hospitals to ensure medical necessity of resource utilization during hospitalization. The concurrent review criteria may be either referred from guidelines or determined by specific clinical attributes such as severity of illness and intensity of services.<sup>3</sup> Both approaches require initial screening which are normally performed by utilization review (or management) nurses (UR or UM nurses) to compare patient's clinical conditions, including both objective data (e.g. blood pressure, body temperature, laboratory and imaging results) and subjective data (e.g. chief complaint, present illness, physical examination) to a set of criteria. The data to be reviewed are from various sources such as Hospital Information System (HIS), peripheral systems, and medical records. In addition to the difficulties in complicated and time-consuming review process, human errors from data oversight have occurred frequently due to data overload.<sup>4</sup> Moreover, the reviewers' performance and accuracy of review results largely depend on the individual's clinical knowledge, cautiousness, and experience. An inexperienced UM nurse might not be able to prioritize work and spend time on low-value tasks, making unnecessary utilization of resources unattended and subsequently causing difficulties among the patient, provider, and payer.

The growth of health insurance industry in Thailand has more than doubled in the past 7 years. Total health insurance premium per annum has continuously increased from 43.4 billion in 2012 to 91.5 billion baht in 2019.<sup>5</sup> This indicates the increasing demand of human resources for both insurance companies and hospitals to handle claim processes and concurrent review. In the US, automated processing and assessment systems have been invented to reduce administrative burden, paperwork, cost, and also to support decision making process.<sup>6-8</sup> The feasibility of Natural Language Processing (NLP) for narrative medical record processing was explored as early as 1981 by Hirschman who used NLP to analyze discharge summary, implemented evaluation criteria, generated evaluation results, and compared to those from physician reviewer.<sup>9</sup>

Currently, extensive AI-enhanced solutions have been offered by many IT companies to lessen UR efforts, such as Case Advisor Services by Optum360° and CORTEX®: The Precision Utilization Management Platform by XSOLIS.<sup>10,11</sup> However, those may not be suitably applied to the insurance industry in Thailand due to different language of medical record documentation and different contextual factors, especially the high proportion of inappropriate "simple diseases" admissions.<sup>12</sup> These diseases can normally be treated in outpatient department. Moreover, Thai Natural Language Processing has continuously

evolved for decades but is not yet fully developed due to words and sentences complexity of Thai language.<sup>13</sup> Hence, new methods or innovative approaches should be introduced to address these problems.

Research and development on machine learning and deep learning in NLP tasks have advanced dramatically. Many progressed machine learning models were developed from the neural network and the variances of it, and each of them has different characteristics and applications. Some research has been conducted on implementing Convolutional Neural Network (CNN) in a sentence classification task and achieved excellent results, even when the training data came from many sources.<sup>14</sup> Moreover, there is research that implements Recurrent Convolutional Neural Network, a hybrid neural network in a sentence classification task, by integrating CNN and Long-Short Term Memory (LSTM) so that the model can capture contextual information.<sup>15,16</sup> As for the problem of Thai language in NLP, PyThaiNLP was developed specifically for this issue.<sup>17</sup> It can segment and tokenize Thai words efficiently and was selected as a tool in our task.

Although a health information entry in a structured format tremendously supports subsequent computer processing, narrative language and semi-structured data may have superior advantages of comprehensiveness, clinician's thought process, and fine structure representation.<sup>18</sup>

As a result, Bangkok Dusit Medical Services, Plc., (BDMS), the largest private hospital network in Thailand, designed and internally developed "BDMS Utilization Review Technology (BURT)" as a decision support application to detect inappropriate hospitalization using natural language processing and a rule-based approach. The implementation of BURT will significantly reduce the assessment processing time performed by UM nurses, since it immediately captures all necessary data at their fingertips. The precision of review results is expected to be higher than those performed by inexperienced UM nurses who occasionally missed crucial data. Moreover, BURT should help expediting the claim process since it will help prioritizing cases that require attention and prompt actions, such as to improve quality of clinical information provided to payers or to clearly communicate to patients regarding unnecessary hospitalization requests.

Among several techniques of machine learning, Neural Network processing is very popular due to its broad learning abilities. As a result, various models have been applied. Convolution Neural Network (CNN) is an advanced form of neural network that can group sentences with a great variety of semantic categories and has been tested with four types of the CNN's default management model.<sup>14</sup> With the Recurrent Convolution Neural Network and Hybrid Neural Network research that apply the Long-Shot Term Memory to CNN to increase the ability to learn, CNN can learn the continuation

of various sentence styles<sup>15,16</sup> and be more efficient with different structure.<sup>19</sup> However, due to the complexity of the internal structure of CNN, which is caused by the application of hierarchical mathematical procedures, there are mysteries that many researchers are trying to understand and interpret. To be able to properly provide information and increase work efficiency<sup>20</sup> from these successful researches, it is part of the approach and inspiration to apply CNN to Natural Language Processing (NLP).

Using the CNN in the BERT program allows us to abstract and interpret free-text data in medical records to see if a word or sentence meets certain criteria. The system has a set algorithm based on data from both NLP and rule-based approach.

## Materials and Methods

### *Establishment of admission criteria based on current medical standard of practice*

To make the tool worthwhile and suitably applicable for the Thai context, it should be able to address a common problem of unnecessary simple disease admissions. Since the definition of simple disease is not universally agreed upon and there is no gold standard of admission guideline for simple disease, we appointed an expert panel of UM physicians including 5 physicians from different specialties and different hospitals. All of them have more than 10 years of medical practice experience and over 2 years of UM physician experience. The panel studied and compiled criteria from several sources including up-to-date, guidelines from The Royal College of Physicians of Thailand, National Clinical Practice, etc.<sup>21-26</sup> and subsequently developed a comprehensive guideline for appropriate hospitalization. The details of 48 criteria variables were defined as A1 to F1 and can be seen in Appendix A.

### *Development of BDMS Utilization Review Technology (BERT) Web Application*

We designed and developed BERT web application to evaluate appropriateness of hospitalization. It consists of two main components, including Prediction Model (AI and Rule-based) to analyze appropriateness of hospitalization based on admission criteria, and Display Application to present predicted result of each case.

### *Development of prediction model (BERT version 1.0)*

We acquired data from a hospital in BDMS network where electronic medical records (EMR) have been completely implemented. Anonymized data of insured-patients admitted from January to May 2019 (1,000 cases) were included to build

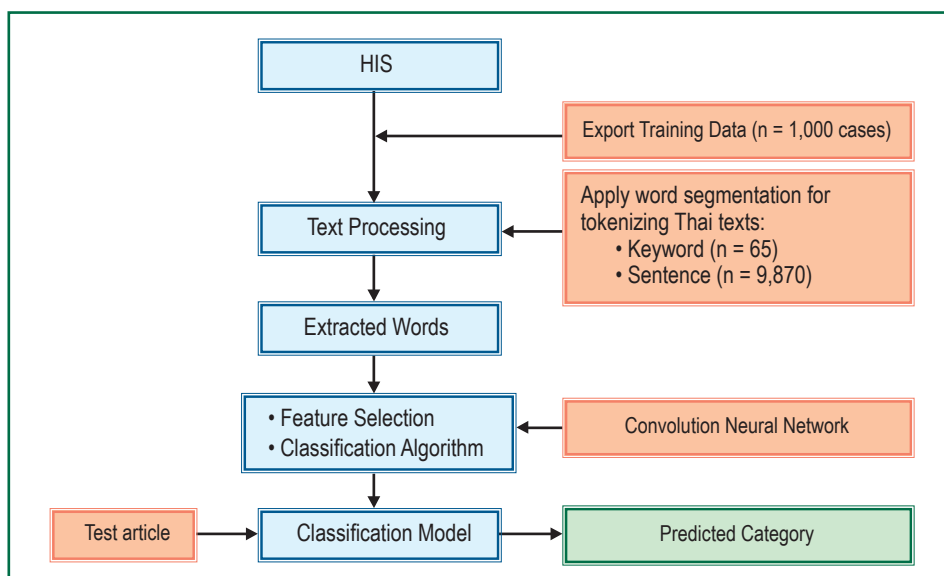
a prediction model. (Demographic information can be found in Appendix B). This dataset was used for the development of prediction model, both AI (Natural Language Classification Model) part and rule-based prediction part. Free-text documented dataset from History of illness, Physical examination, and from some physicians' order notes, contributed to development of Natural Language Classification Model.

From admission criteria, we defined 65 terms for NLP training, for example "cannot walk", "weakness in arms and legs" "mental status change", "swollen face", "swollen eyelid", "very fatigued" etc. for B1 criterion - General Appearance. These were sorted from the abovementioned dataset, which made altogether 9,870 sentences. Data were randomly selected into two sets of 80% (7,896 sentences) for model training and 20% (1,974 sentences) for model testing. From all text data, sentences were split into words for training with PyThaiNLP (Appendix C), which is a Python library to truncate words.<sup>17</sup> A word segmentation technique was deployed in both training step and testing step in our research. A window size of 7 was used (7 before and 7 after the context word). The expert team made labels of each phrase to indicate whether those phrases corresponded with desired criteria, using the binary codes of 1 for corresponded and 0 for not corresponded. Again, PythaiNLP package can facilitate transformation of words encoding into randomly vectorized numerical data for CNN training in the next step.

Labeled data and data from word tokenization was trained using CNN model, which consists of 2 layers, 3x1 convolutional layer and dense layer. For dense layer, the number of neural network nodes depended on the number and complexity of the training sentences, making the overall model consists of multiple models.

The trained CNN model was implemented for each single NLP criterion and provided a predictive result from all medical sentence input. Therefore, each NLP criterion will be classified as "Met" or "Not Met". Flow chart of Natural Language Classification model of this project is shown in Figure 1. The performance using testing data as an input achieved 95% accuracy.

A rule-based module was set up with conditional values taken from vital signs, laboratory results, and computerized physician order entry (CPOE) (Table 1). The criteria rule was based on a comprehensive guideline for appropriate hospitalization compiled by UM Physician Panel. This combined AI and rule-based algorithm was then able to evaluate appropriateness of admission and show results as "Appropriate hospitalization" or "Inappropriate hospitalization".



**Figure 1:** Flow Chart of Natural Language Classification Model (BURT 1.0)

**Table 1:** Dependent variables and clinical data sources used by BURT predictive algorithm

Admission criterion	Classification Technique	OPD & ER record	T.P.R. record	Pain record	Lab	Imaging	EKG	Order sheet	MAR sheet	Order item	Operative note
<b>A History of illness</b>											
A1 Neurological/Cardiovascular	NLP	√									
A2 Cardiovascular/ Lower respiratory problems	NLP	√									
A3 Chest pain, Suspected ACS	NLP	√									
A4 Unable to eat /Dehydration problems	NLP	√									
A5 Prolong illness	NLP	√									
<b>B Physical Examination</b>											
B1 General Appearance	NLP	√									
B2 Temperature (Non RTI)	NLP&CON	√	√								
B3 High Blood Pressure	CON	√	√								
B4 Low Blood Pressure	NLP&CON	√	√								
B5 Pulse	CON	√	√								
B6 Anemia	NLP	√									
B7 Dehydration problems	NLP	√									
B8 Neurological problems	NLP	√									
B9 Sign of shock	NLP	√									
B10 Respiratory Rate (Non RTI)	NLP&CON	√	√								
B11 Oxygen Saturation	CON	√	√								
B12 Lower respiratory problems (Abnormal breath sounds), Cardiovascular problems	NLP	√									
B13 Pain score	CON		√	√							
B14 Surgical abdomen	NLP	√									
B15 Temperature (RTI)	NLP&CON	√	√								
B16 Respiratory (RTI)	NLP&CON	√	√								
B17 Temperature (RTI + Non RTI)	NLP&CON	√	√								
<b>C Investigation</b>											
C1 White Blood Cell (WBC)	CON				√						
C2 Platelets	CON				√						
C3 Hemoglobin (Hb)	CON				√						
C4 Hematocrit (Hct)	CON				√						

Admission criterion		Classification Technique	OPD & ER record	T.P.R. record	Pain record	Lab	Imaging	EKG	Order sheet	MAR sheet	Order item	Operative note
C5	Serum Sodium(Na+)	CON				√						
C6	Serum Potassium(K+)	CON				√						
C7	Bicarbonate (Total CO2)	CON				√						
C8	Creatinine	CON				√						
C9	Glucose	CON				√						
C10	Urine specific gravity	CON				√						
C11	Urine ketone	CON				√						
C12	Troponin-T	CON				√						
C13	Troponin- I	CON				√						
C14	ECG / EKG	NLP & CON						√				
C15	• CT / MRI / MRV /MRA (Brain, Chest, Abdomen) • Film Chest x-ray	NLP & CON					√					
<b>D Management</b>												
D1	Observation / Monitoring q 4 hrs or more	NLP							√			
D2	Oxygen supplement	CON							√		√	
D3	IV fluid maintenance or more ml/ hour	CON								√	√	
D4	Bronchodilator NB at least 3 in 24 hours or more	CON								√		
D5	IV or IM medication at least 2 in 24 hours or more	CON									√	
D6	Blood transfusion or blood components	CON									√	
<b>E Procedures</b>												
E1	Procedure under Spinal block	CON										√
E2	Non-minor procedure under General anesthesia	CON										√
E3	Minor procedure under General anesthesia	CON										√
E4	Cardiac procedures	CON									√	
<b>F Age</b>												
F1	Age <1 and > 75	CON	√									

\*NLP = Natural Language Processing, CON = Value Condition or Data Condition (Rule-based)

### Fine-tuning process

Data from 500 cases, which were the subset of 1,000 cases (Appendix B), were randomly selected and comprehensively studied by the UM physician expert panel. A series of collaborative discussions among UM Physicians gearing towards final agreement for each case was made from a majority vote. The mutual goal was threefold: to improve NLP accuracy, to evaluate appropriateness of admission criteria algorithm setup, and to evaluate the effectiveness of automated application of the criteria. We aimed to make BURT capable of analyzing appropriateness of hospitalization in both simple diseases and non-simple diseases with various levels of severity. During the fine-tuning process, we increased keywords for NLP training (from 65 to 79 terms), and increased NLP training and testing dataset (from 9,870 to 77,707

sentences). The anonymized data of insured-patients admitted between December 2017 and October 2019 (22,020 cases) were included. The rule-based model was also enhanced from condition-based to scoring-based in order to classify different levels of severity. After the fine-tuning process, tremendous improvement was made from our first to the latest version of the tool, and can be described as BURT 1.0 and BURT 1.1 respectively. Examples of major improvement are explained in Appendix D.

After several consecutive cycles of NLP training and testing, we achieved great results of NLP model with significantly increasing accuracy, precision, recall, and F1-score. (Table 2). The definition of each matrices can be seen in Figure 2.



$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

$$Precision = \frac{\text{positive sentences correctly identify}}{(\text{positive sentences correctly identify} + \text{positive sentences incorrectly identify})}$$

$$Recall = \frac{\text{positive sentences correctly identify}}{(\text{positive sentences correctly identify} + \text{negative sentences incorrectly identify})}$$

$$F1\text{-score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

**Figure 2:** The definition of evaluation matrices.

**Table 2:** Confusion matrix for NLP predicted model in BURT 1.0 and BURT 1.1

Predicted value	BURT 1.0	Actual value		Accuracy
	n = 1,974 sentences	Positive	Negative	Precision
	Positive	1229	70	Recall
	Negative	88	587	F1 score
Predicted value	BURT 1.1	Actual value		Accuracy
	n = 15,507 sentences	Positive	Negative	Precision
	Positive	8125	64	Recall
	Negative	68	7250	F1 score

The agreement between BURT 1.1 and UM Physician Panel also increased significantly from 70% to 85% (Table 3). We intended to minimize the number of false positive (false appropriate), since it will make UM nurse inadvertently rely upon the predicted results and overlook any issues that require a necessary action. We then thoroughly reviewed all the 14 false positive (false appropriate) cases and found out that the agreements on inappropriate hospitalization among UM Physician Panel were not unanimous. Ultimately, BURT 1.1 showed very promising predicted results. However, given that advanced medical technology and treatments are continuously being evolved, we will continue to align our tool accordingly.

### Display Application

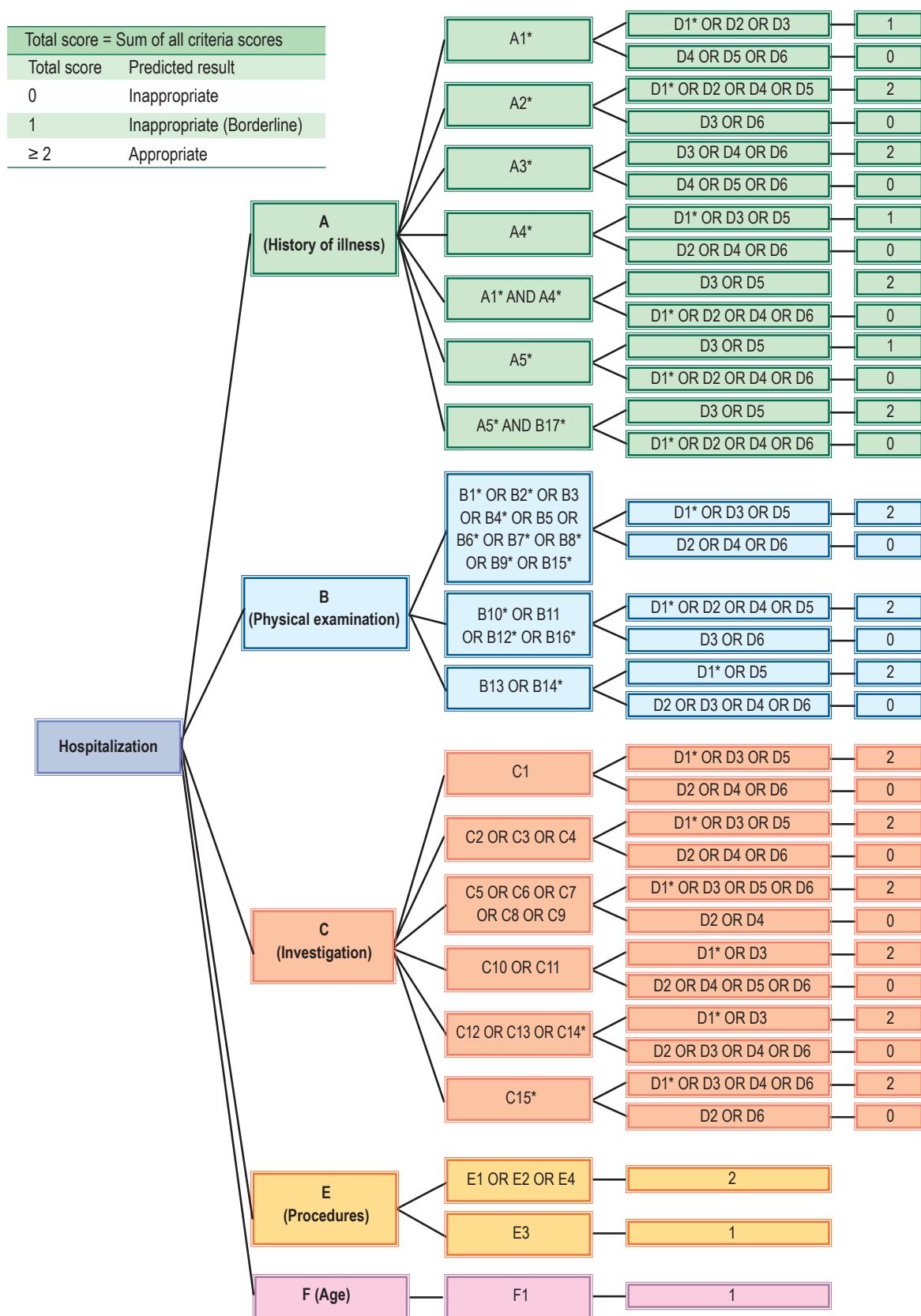
The predicted results are displayed on web application. The dependent variables of appropriate hospitalization decision were categorical and have three possible results: appropriate admission, borderline inappropriate admission, and inappropriate admission. (Figure 3 and Figure 4)

- Inappropriate admission (Total score = 0) means admission criteria are not satisfied. The case will be shown in red color, displayed on the top, and should be given first priority by UM nurse.
- Borderline inappropriate admission (Total score = 1) means admission criteria are partially satisfied, but still require manual review by UM nurse. The case will be shown in yellow color, displayed beneath the red ones, and should be given second priority by UM nurse.
- Appropriate admission (Total score ≥ 2) means admission criteria are strongly fulfilled. The case will be shown in green color and displayed at the bottom. UM nurse can simply disregard this category unless for educational purposes.

The prediction outcome and detailed information for each criterion will be displayed on the screen so that users can take suitable further action, or to notice the weakness of algorithm, or even to make recommendation and give feedback to the application administrator.<sup>27</sup>


**Table 3:** Confusion matrix for appropriate vs inappropriate hospitalization predicted by 2 versions of BURT

		UM Physician Panel Consensus	
		Appropriate (Positive)	Inappropriate (Negative)
BURT Predicted Result	BURT 1.0	Appropriate (Positive)	301
	BURT 1.1	Inappropriate (Negative)	67
		Appropriate (Positive)	81
		Inappropriate (Negative)	51
	BURT 1.1	Appropriate (Positive)	339
		Inappropriate (Negative)	14
	BURT 1.1	Appropriate (Positive)	61
		Inappropriate (Negative)	86



Remark: (\*) means criteria that NLP technique is implemented along with rule-based. Criteria with no (\*) means only rule-based is implemented. Definitions and criteria examples of each code can be seen in Appendix A.

**Figure 3:** Admission criteria scoring system implemented in BURT 1.1: NLP and rule-based algorithm


**BDMS Utilization Review Technology**

Claim
Admin
User

Hospital Site :


Date Filtering
EN Searching

Date Criteria :
Start Date :
End Date :

Admit Date
08/06/2020 13:00
09/06/2020 23:30

Search
Review History Search
Export Excel


No.1
EN :
Gender : XXXX
Age : XXXX
Visit Date/Time : XXXXXXXX
First Seen Doctor : XXXXXXXX



History of illness		Physical Examination		Investigation		Exclusion
Main Condition	Management	Main Condition	Management	Main Condition	Management	Main Condition

Inappropriate (0)


No.102
EN :
Gender : XXXX
Age : XXXX
Visit Date/Time : XXXXXXXX
First Seen Doctor : XXXXXXXX



History of illness		Physical Examination		Investigation		Exclusion
Main Condition	Management	Main Condition	Management	Main Condition	Management	Main Condition

Borderline (1)

No.149
EN :
Gender : XXXX
Age : XXXX
Visit Date/Time : XXXXXXXX
First Seen Doctor : XXXXXXXX



History of illness		Physical Examination		Investigation		Exclusion
Main Condition	Management	Main Condition	Management	Main Condition	Management	Main Condition

Appropriate (2)

Figure 4: Screen shot of BURT1.1 displaying prediction results on hospitalization appropriateness.

## Results

### BURT1.1 performance evaluation

We evaluated the performance of BURT1.1 in assessing inappropriate hospitalization by comparing its results to those manually performed by UM nurses. Using the identical 300 cases, we compared the review results from three sources, including BURT1.1, concurrent review data collection from the participating hospital, and retrospective review report from Bangkok Hospital Headquarters. Firstly, we randomly selected 300 cases from 500 cases that had been thoroughly reviewed by UM Physician Panel and made a consensus on appropriateness of hospitalization. Secondly, we searched for the review results of those 300 cases that were manually reviewed by UM nurses in the participating hospital and concurrently reviewed during patient stay. Thirdly, we retrieved detailed clinical data of 300 cases from important sources that

required by BURT1.1 algorithm and summarized them in a ready-to-review excel form, and assigned to UM nurses at Bangkok Hospital Headquarters. Three UM nurses with different levels of experience, including 1 in-charge UM nurse and 2 operational-level UM nurses were specifically selected to perform secondary analysis of retrospective data. The criteria set was clearly explained prior to the review process in order to reduce variance. Rate of agreements on detection of inappropriate hospitalization between UM Physician Panel and BURT1.1, UM Physician Panel and concurrent review report from the network hospital, and UM Physician Panel and retrospective review report from Bangkok Hospital Headquarter were then compared (Table 4).



**Table 4:** Rates of agreement between UM Physician Panel and different methods of review, namely BURT1.1, concurrent review by UM nurses at network hospital, and retrospective review by UM nurses at Bangkok Hospital Headquarter (BHQ).

Method of review	UM Physician Panel Consensus		Rate of agreement
	Agree	Disagree	
Concurrent review summary report from network hospital	218	82	72.67%
Retrospective review by UM Nurse 1 (Operational level) from BHQ	228	72	76.00%
Retrospective review by UM Nurse 2 (Operational level) from BHQ	232	68	77.33%
Retrospective review by UM Nurse 3 (In-charge level) from BHQ	237	63	79.00%
BURT1.1	258	42	86.00%

Overall, BURT1.1 showed a favorable outcome with the highest rate of agreement (86.00%) among all the methods. An in-charge UM nurse and two operational-level UM nurses from Bangkok Hospital Headquarters had 79%, 77.33%, and 76.00% respectively. Network hospital showed the lowest rate of agreement (72.67%), presumably from excessive workload and a laboriously manual review process that all the scattered sources of information might easily lead to data oversight.<sup>28</sup> The performance evaluation results demonstrated that BURT1.1 can support utilization review process in detecting inappropriate admission and leading to an appropriate action that should be taken by a UM nurse.

#### *BURT1.1 prediction output and insurance claim approval*

We retrospectively evaluated the insurance claim adjudication decision made by insurance companies for those 300 cases (Table 5). The claim approval apparently aligned with the BURT1.1 prediction outputs. As we prioritized precision over recall, 99% of precision represented a satisfied prediction

outcome. There were three cases that BURT1.1 predicted output showed appropriate, but the claims were denied by insurance companies. We thoroughly reviewed the reason of claim denial for these three particular cases and found out that they were non-clinical issue of suspected preexisting conditions. On the other hand, in Thailand, the claim denial rate for inappropriate hospitalizations was relatively low due to the fact that various factors are usually incorporated in claim adjudication process especially special business condition that can overrule medical necessity and insurance policy details. However, BURT1.1 was not designed to predict claim approval decision and will definitely assist UM nurses in detecting medically inappropriate admission, inappropriate documentation, and patient's request for unnecessary admission. One case rejected by an insurance company due to unnecessary admission, for instance, BURT1.1 would have displayed the unmet criterion which was actually from incomplete documentation. Feedback to physician should have been provided immediately to improve medical record documentation.

**Table 5:** Confusion matrix of BURT1.1 prediction output and insurance claim approval

Insurance claim	BURT1.1 prediction output		
	Approve	Not approve	
Appropriate	203	3	Accuracy 0.69
Inappropriate	89	5	Precision 0.99
			Recall 0.70
			F1 score 0.82

#### *Time and cost -saving utilization review process*

Processing time of conventional admission review that has been manually performed by UM nurses, varies between 10-15 minutes per case, depending on UM nurse experience, medical knowledge, and competency. Within 0.59 seconds, BURT1.1 can provide a prediction result output, this allows UM nurses to perform more valuable tasks, increases productivity, and reduces operating costs. For example, for the total of 300 cases in our BURT1.1 performance evaluation, UM nurses could have dismissed 206 (69%) cases, and directly focus on only 94 cases.

#### **Discussion**

The significance of our study is that the predictive algorithm development process has been exhaustively refined by a UM physician expert panel. Additionally, the combination of Thai natural language processing and a rule-based model, which runs on web application, makes it highly applicable for other hospital settings. However, pre-required electronic medical record (EMR) and computerized physician order entry (CPOE) are unavoidable, since the tool needs electronic records for rule-based approach.

A limitation of this study is that the data is from only one hospital, so overfitting could be a problem that we would expect to encounter due to specific practice patterns and clinical cultures of each hospital. Similar to other previous studies, no clinical NLP algorithm is completely accurate and we also found the same challenges of missing information or medical narrative explaining severity of illness, not in a guideline terminology format but implicitly stated in documents.<sup>27, 29</sup>

BURT1.1 will be implemented as an automatic routine screening tool in a pilot hospital in the BDMS network. It will capture data directly from the Hospital Information System (HIS) and related peripheral systems, and perform hospitalization appropriateness evaluation. However, further work needs to be contributed by both developers and users to make the tool functioning more accurately and to fix the overfitting problem. Under certain circumstances, disagreement between BURT1.1 and human expert may occur and should be discussed with an application administrator. Potential improvement would be to refine the precision of NLP and measure sensitivity and specificity of the tool. Advanced scoring system for severity of illness, including patient's chief complaint, clinical signs, and symptoms, should be established to provide probabilistic outputs in more details. Additional thesaurus of signs and symptoms may be required for extra training. The tool should be directly integrated into BDMS e-Claim system and shared display application in order to perform real-time evaluation. To make the tool easily applicable to analyze evidence-based guideline compliance for other resource utilization, such as appropriateness of diagnostic and therapeutic services, we plan to enhance NLP tasks to be able to extract temporal information, track longitudinal progression of an illness,<sup>30</sup> and transform unstructured data to structured data.<sup>31</sup>

## Conclusion

Through several collaborative discussions among Utilization Management experts with available electronic

health record (EHR) data, we developed an AI enhanced tool using Natural Language Processing and Rule Based algorithms, namely BDMS Utilization Review Technology (BURT) and have improved version as BURT1.1. This BURT1.1 should be effectively implemented as an automatic daily screening tool for inappropriate hospitalization, which is a first level review process of utilization management. It can immediately identify records at high risk of inappropriate hospitalization that require further manual review by UM nurse, provide feedback to attending physician on the completeness and quality of documentation, as well as notifying the case to be reviewed by UM physician. BURT1.1 will also be beneficial to UM nurses to develop their knowledge and professional judgement, since it will clearly show on the display screen how patient's data satisfies each specific criterion. Ultimately, the use of BURT1.1 would increase UM efficiency, expedite the claim process due to more complete data submissions, reduce health care costs from unnecessary hospitalization, and reduce claim denials.

## Acknowledgements

We thank members of UM expert panel, (Piemchok Banomyong, MD, Suwapat Dewan, MD, Chatchai Charoensri, MD, Wittaya Konngam, MD, Naruenart Koovimon, MD), UM nurses of both participating hospital and Bangkok Hospital Headquarters. Behind every step, we received technical support from hospital IT staff, GreenLine Synergy teams and Wuttichai Luangruangrong, PhD. We thank them for that. Finally, this study would never been completed without management supports of Mrs. Narumol Noi-am, Senior Executive Vice President of BDMS, Trin Charumilind, MD-Chief of Doctors, Matinee Maipang, MD-hospital director of Bangkok Hospital Headquarters, and hospital director of the participating hospital.

## References

1. Dijk CE, van Berg B, van den Verheij RA, et al. Moral Hazard and supplier-induced demand: empirical evidence in general practice. *Health Econ* 2013;22(3):340-52.
2. Folland S, Goodman AC, Stano M. Demand and Supply of Health Insurance. *The Economics of Health and Health Care* 2013; 8<sup>th</sup> ed.:148.
3. Thomas M, Wickizer I and Daniel Lessler. Utilization Management: Issues, Effects, and Future Prospects. *Ann Rev Public Health* 2002;23:233-54.
4. Nelson BD, Gardner RM. Decision support for concurrent utilization review using a HELP-embedded expert system. *Proc Annu Symp Comput Appl Med Care* 1993:176-82.
5. Office of Insurance Commission. Thailand. Annual report 2012-2019.
6. John William Richards JR. Inventor: Systems and methods for automated processing and assessment of an insurance disclosure via a network. United States patent US20070088579A1.2007 Apr 19.
7. Kirsh WD, Kramer PM, King JT, inventor: System and method for standardized and automated appeals process. United States patent US8204765B2. 2012 Jun 19.
8. Steffen Hehner, Boris Kors, Manuela Martin, Elke Uhrmann-Klingen, Jack Waldron; Copyright © McKinsey & Company. Artificial intelligence in health insurance. Smart claims management with self-learning software. *Healthcare*. September 2017:1-11.

9. Hirschman L, Story G, Marsh E, et al. Experiment in automated health care evaluation from narrative medical records. *Comp Biomed Research* 1981;14(5):447-63
10. Optum360, LLC. Case Advisor Services AI-Powered Physician Advisor Solutions.(Accessed May 15, 2020, at <https://www.optum.com/resources/library/ai-powered-services.html>.)
11. XSOLIS. The Precision Utilization Management Platform. (Accessed May 15,2020, at <https://www.xsolis.com/solutions/provider-overview>)
12. Duangnet C, Panyasorn J, Phengpinit N, et al. Rational Identification of Simple Disease Cases in Bangkok Dusit Medical Services Hospitals using Relative Weight and Case Mixed Index. *BKK Med J* 2019;15(2):130-9.
13. Tapsai C, Meesad P, Unger H. An Overview on the Development of Thai Natural Language Processing. *Inform Techno J* 2019;15(2):45-52.
14. Kim Y. Convolutional neural networks for sentence classification. New York: arXiv, Cornell University and Simon Foundation; 2014 Sep 3. Report No.: arXiv:1408.5882.
15. Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification. The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15); 2015 January 25–30; Texas, USA. Austin: AAAI Press; 2015.
16. Zhou Z, Zhu X, He Z, et al. Question classification based on hybrid neural networks. In 2016 4th International Conference on Electrical & Electronics Engineering and Computer Science (ICEECS 2016). Atlantis Press, 2016b.
17. PyThaiNLP team, © Copyright 2017-2019, PyThaiNLP (Apache Software License 2.0) (Accessed May 15,2020, at <https://www.thainlp.org/pythainlp/docs/dev/index.html>).
18. Johnson SB, Bakken S, Dine D, et al. An Electronic Health Record Based on Structured Narrative. *J Am Med Inform Assoc* 2008;15(1):54-64.
19. Rie J, Tong Z. Deep pyramid convolutional neural networks for text categorization. In ACL, 2017.
20. Jacovi A, Shalom OS, Goldberg Y. Understanding convolutional neural networks for text classification. New York: arXiv, Cornell University and Simon Foundation; 2020 Apr 27. Report No. arXiv:1809.08037.
21. Scottish Intercollegiate Guidelines Network. Bronchiolitis in children, a national clinical guideline. 2006;91:1-40
22. Ministry of Public Health. Clinical Practice Guideline for Influenza. 3 ed 27 September 2011:1-15.
23. Haffner J, Parsons D. Clinical Presentation Guidelines for Use in the Minor Injury Units. 2011:1-65
24. Ministry of Public Health - Lebanon. Medical necessity for Inpatient admission: Banner Health System Observation Group. 2013:1-4
25. PHO Alliance inc/Primary Health Alliance inc. Primary Options for Acute Care(POAC) Clinical Guideline: Acute Adult Dehydration 2015:1-3
26. Thai Society of Pediatric Gastroenterology and Hepatology. Clinical Practice Guideline for Acute Diarrhea in Children. 2019:1-56
27. Velupillai S, Suominen H, Liakata M, et al. Using Clinical Natural Language Processing for Health Outcomes Research: Overview and Actionable Suggestions for Future Advances. *J Biomed Inform* 2018;88:11-9.
28. Nelson BD, Gardner RM, Hedrick G, et al. Computerized decision support for concurrent utilization review using the HELP system. *J Am Med Inform Assoc* 1994; 1(4): 339-52.
29. Lenert LA, Tovar M. Automated Linkage of Free-Text Descriptions of Patients With a Practice Guideline. *Proc Annu Symp Comput Appl Med Care* 1993;274-8.
30. Metfessel BA. An Automated Tool for an Analysis of Compliance to Evidence-based Clinical Guidelines. *Stud Health Technol Inform* 2001;84(Pt 1):226-30.
31. Friedman C, Rindflesch TC, Corn M. Natural Language Processing: State of the Art and Prospects for Significant Progress, a Workshop Sponsored by the National Library of Medicine. *J Biomed Inform* 2013;46(5):765-73.

## Appendix A

Examples of Hospitalization appropriateness criteria as implemented in BURT. History of illness, physical examination, laboratory and imaging results, and relevant treatment plan, are altogether taken into account in order to fulfill the admission criteria.

Variables	Examples of criteria
<b>A History of illness</b>	
A1 Neurological/ Cardiovascular	Near syncope, Micturition syncope
A2 Cardiovascular/ Lower respiratory problems	Progressive dyspnea, dyspnea at rest
A3 Chest pain, Suspected ACS	Unstable angina
A4 Unable to eat/ Dehydration problems	Vomiting more than 2 times
A5 Prolonged illness	Prolonged illness $\geq 3$ days with Temp $> 38.3^{\circ}\text{C}$
<b>B Physical Examination</b>	
B1 General Appearance	Angioedema Unable to move Mental status change
B2 Temperature (Non RTI)	Temp $> 38.3^{\circ}\text{C}$
B3 High Blood Pressure	SBP $> 185$ mmHg or DBP $> 120$ mmHg (Age $\geq 9$ years)
B4 Low Blood Pressure	SBP $< 85$ mmHg or DBP $< 50$ mmHg (Age $\geq 9$ years)
B5 Pulse (Non RTI)	Pulse $< 40$ , $> 120$ / min , $\geq 100$ (Age $\geq 65$ years)
B6 Anemia	Pale skin
B7 Dehydration problems	Sunken eye ball
B8 Neurological problems	New abnormal / Focal signs neurological exam
B9 Sign of shock	Cold/ Clammy extremities, Capillary refill time (CRT $> 2$ sec) Faint pulse
B10 Respiratory Rate (Non RTI)	RR $> 30$ /min RR $> 50$ /min (Age $\leq 12$ months) RR $> 40$ /min (Age $> 12$ months - 3 years)
B11 Oxygen Saturation	$< 95\%$
B12 Lower respiratory problems (Abnormal breath sounds), Cardiovascular problems	Wheezing/ Poor air entry/ Stridor/ Rhonchi/ Chest retractions
B13 Pain score	Pain score $\geq 8$
B14 Surgical abdomen	Rebound tenderness/ Guarding
B15 Temp (RTI)	Temp $\geq 38.9^{\circ}\text{C}$
B16 Respiratory (RTI)	RR $> 40$ /min (Age $\leq 5$ years) , RR $> 30$ /min (Age $> 5$ years)
B17 Temp (RTI+Non RTI)	Temp $> 38.3^{\circ}\text{C}$
<b>C Investigation</b>	
C1 White Blood Cell (WBC)	$< 3,000$ or $> 12,000$ cell / cu.mm
C2 Platelets	$< 100,000$ or $> 500,000$ cell / cu.mm
C3 Hemoglobin (Hb)	$< 9.5$ or $> 18$ g/dl
C4 Hematocrit (Hct)	$< 29$ or $> 55\%$
C5 Serum Sodium(Na+)	$< 130$ or $> 150$ mmol/ L
C6 Serum Potassium(K+)	$< 3.1$ or $> 6$ mmol/ L
C7 Bicarbonate (Total CO2)	$< 16$ or $> 34$ mmol/ L
C8 Creatinine	$> 1.4$ mg/ dL
C9 Glucose	$< 60$ or $> 300$ mg/ dL with symptomatic
C10 Urine specific gravity	$> 1.025$
C11 Urine ketone	$> 1 +$ mg/dL
C12 Troponin-T	$> 0.013$ ng/mL
C13 Troponin- I	$> 0.001$ ng/mL
C14 ECG / EKG	New onset abnormal
C15 CT/ MR/ MRV/ MRA (Brain, Chest, Abdomen), Film Chest x-ray	CT/ MRI (Hemorrhage, Hematoma) Chest x-ray (Infiltration)

Variables		*20 Minor Procedures
D Management		
D1	Observation / Monitoring q 4 hrs or more	• ESWL: Extracorporeal Shock Wave Lithotripsy
D2	Oxygen supplement	• Coronary Angiogram/ Cardiac Catheterization
D3	IV fluid maintenance or more ml/ hour	• Extra Capsular Cataract Extraction with Intra Ocular Lens
D4	Bronchodilator NB at least 3 in 24 hours or more	• Endoscope
D5	IV or IM medication at least 2 in 24 hours or more	• Sinus Operations
D6	Blood transfusion or blood components	• Injection or Rubber band ligation
E Procedures		• Excision breast mass
E1	Procedure under Spinal block	• Bone biopsy
E2	Non-minor procedure under General anesthesia	• Tissue biopsy
E3	Minor procedure* under General anesthesia	• Ray Amputation/ Toe amputation/ Finger amputation/ Phalange amputation
E4	Cardiac procedures	• Manual reduction
F Age		• Liver Puncture/ Liver Aspiration
F1	Age <1 and > 75 years	• Bone Marrow Aspiration
		• Lumbar Puncture
		• Thoracentesis/ Pleurocentesis/ Thoracic Aspiration/ Thoracic Paracentesis
		• Abdominal Paracentesis/ Abdominal tapping
		• Curettage/ Dilatation & Curettage/ Fractional Curettage
		• Colposcopy/ Loop diathermy
		• Marsupialization of Bartholin's Cyst
		• Gamma knife

## Appendix B

Demographic and clinical characteristics of cases in this study

	Train & Test BURT 1.0 (1,000 cases)	Model fine-tuning ( 500 cases )	Train & Test BURT 1.1 ( 22,020 cases)	Performance evaluation BURT 1.1 ( 300 cases )
Age				
0-19	412 (41.2%)	222 (44.4%)	13,519 (61.4%)	145 (48.3%)
20-39	268 (26.8%)	129 (25.8%)	4,301 (19.5%)	78 (26.0%)
40-59	191 (19.1%)	99 (19.8%)	3,217 (14.6%)	58 (19.3%)
≥60	129 (12.9%)	50 (10.0%)	983 (4.5%)	19 (6.3%)
Sex				
Male	471 (47.1%)	235 (47.0%)	10,731 (48.7%)	146 (48.7%)
Female	529 (52.9%)	265 (53.0%)	11,289 (51.3%)	154 (51.3%)
Disease				
Simple disease	595 (59.5%)	305 (61.0%)	14,626 (66.4%)	198 (66.0%)
Non-simple disease	405 (40.5%)	195 (39.0%)	7,394 (33.6%)	102 (34.0%)
Specialty				
Medicine	749 (74.9%)	384 (76.8%)	19,168 (87.0%)	240 (80.0%)
Surgical	22 (2.2%)	15 (3.0%)	700 (3.2%)	9 (3.0%)
Neurology	35 (3.5%)	27 (5.4%)	533 (2.4%)	12 (4.0%)
Orthopedics	31 (3.1%)	17 (3.4%)	451 (2.0%)	4 (1.3%)
Cardiology	92 (9.2%)	25 (5.0%)	405 (1.8%)	19 (6.3%)
Ob-Gyn	37 (3.7%)	16 (3.2%)	363 (1.6%)	3 (1.0%)
Urology	21 (2.1%)	9 (1.8%)	182 (0.8%)	6 (2.0%)
Oncology	5 (0.5%)	5 (1.0%)	121 (0.5%)	5 (1.7%)
Ophthalmology	8 (0.8%)	2 (0.4%)	97 (0.4%)	2 (0.7%)
Length of stay (day)				
1-2	700 (70%)	355 (71.0%)	11,795 (53.6%)	215 (71.7%)
3-4	242 (24.2%)	117 (23.4%)	7,498 (34.1%)	68 (22.7%)
≥ 5	58 (5.8%)	28 (5.6%)	2,727 (12.4%)	17 (5.6%)



## Appendix C

Thai word segmentation challenges and solutions.

Thai is an isolating language. There are no clearly defined boundaries of words and sentences.

For example, “ตากลม” can be segmented to “ตา-กลม” and “ตา-ลม”, depending on its context.

Thai language divides differently, meaning changes “ตากลม” → “ตา-กลม” (round eye) or “ตา-ลม” (air dying)

PyThaiNLP is a Python library which includes many Thai text processing techniques, including Thai word segmentation. PyThaiNLP package was then implemented to solve the problem of converting sentences into words.

For example: Thai language does not have word breaks → “ผู้ป่วยมีอาการแน่นหน้าอกหายใจไม่ออก”

There are no gaps in word breaks “patienthas-symptom-chest-pain-be-unable-to-breathe”

The sentence is processed and segmented by PyThaiNLP into:

Thai language can identify words correctly → “ผู้ป่วย-มี-อาการ-แน่นหน้าอก-หายใจ-ไม่-ออก”

“patient-has-symptom-chest pain-be unable to breathe”

## Appendix D

Examples of fine-tuning process from BURT 1.0 to BURT 1.1.

### *Clinical criteria improvement*

- Recategorization of 21 minor procedures (as defined by Office of Insurance Commission) performed under general anesthesia to “Inappropriate hospitalization” except laparoscopic procedures.
- Added code A5 (Prolonged illness) to address history of outpatient treatment failure.
- Added code E4 (Cardiac procedures) to include procedures under local anesthesia eg. Angiography and Percutaneous Coronary Intervention etc.

### *NLP improvement*

- Refined accuracy of each word index, specifically focus on the ones with less than 90% accuracy.
- Increased NLP training dataset, especially for the low-volume word indexes.
- Added NLP training dataset to include prolonged illness.
- Added automated spelling correction function.

### *Classification engine improvement*

- Corrected rule-based algorithm that was erroneously setup.
- Enhanced rule-based model, from condition-based to scoring-based, in order to classify different levels of severity.
- Added classification of “Borderline inappropriate hospitalization”
- Revised scoring criteria of age from “If age < 1 or > 85, then score = 2” to “If age < 1 or > 75, then score = 1”
- Added scoring criteria of combined A1 and A4 (Dehydration problems and Neurological & Cardiovascular problems) to appropriately address higher severity of illness.