Development of an Artificial Intelligence Model for Prediction of Dry Weight in Chronic Hemodialysis Patients and Assessment of its Accuracy Compared to Standard Bioelectrical Impedance Analysis

Mr. Nataphut Boonvisuth

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Medicine
Department of Medicine
FACULTY OF MEDICINE
Chulalongkorn University
Academic Year 2022

การพัฒนาโมเดลด้วยปัญญาประดิษฐ์เพื่อทำนายน้ำหนักแห้งในผู้ป่วยไตวายเรื้อรังระยะสุดท้ายที่ได้รับการฟอกเลือดด้วยเครื่องไตเทียม และการประเมินความแม่นยำของโมเดลโดยเปรียบเทียบกับน้ำหนักแห้งที่ได้จากการวิเคราะห์องค์ประกอบของร่างกายจากความต้านทานของกระแสไฟฟ้า

นายณฐพุฒิ บุญวิสุทธิ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาอายุรศาสตร์ ภาควิชาอายุรศาสตร์
คณะแพทยศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

Thesis Title	Development of an Artificial Intelligence Model for Prediction of Dry Weight in Chronic Hemodialysis Patients and Assessment of its Accuracy Compared to Standard Bioelectrical Impedance Analysis
By	Mr. Nataphut Boonvisuth
Field of Study	Medicine
Thesis Advisor	Professor KHAJOHN TIRANATHANAGUL, M.D.
Thesis Co Advisor	SIRA SRISWASDI, Ph.D.

Accepted by the FACULTY OF MEDICINE, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

..... Dean of the FACULTY OF MEDICINE
(Associate Professor CHANCHAI SITTIPUNT, M.D.)

THESIS COMMITTEE

..... Chairman
(Professor PONLAPAT ROJNUCKARIN, M.D., Ph.D.)

..... Thesis Advisor
(Professor KHAJOHN TIRANATHANAGUL, M.D.)

..... Thesis Co-Advisor
(SIRA SRISWASDI, Ph.D.)

..... Examiner
(Nicha Somlaw, M.D.)

..... External Examiner
(Capt. SUREE YOOWANNAKUL, M.D.)

ณฐพุฒิ บุญวิสุทธิ์ : การพัฒนาโมเดลด้วยปัญญาประดิษฐ์เพื่อทำนายน้ำหนักแห้งในผู้ป่วยไตวายเรื้อรังระยะสุดท้ายที่ได้รับการฟอกเลือดด้วยเครื่องไตเทียม และการประเมินความแม่นยำของโมเดลโดยเปรียบเทียบกับน้ำหนักแห้งที่ได้จากการวิเคราะห์องค์ประกอบของร่างกายจากความต้านทานของกระแสไฟฟ้า. (Development of an Artificial Intelligence Model for Prediction of Dry Weight in Chronic Hemodialysis Patients and Assessment of its Accuracy Compared to Standard Bioelectrical Impedance Analysis) อ.ที่ปรึกษาหลัก : ศ. นพ.ขจร ตีรณธนากุล, อ.ที่ปรึกษาร่วม : ดร.สิระ ศรีสวัสดิ์

ความสำคัญและที่มาของปัญหา: การกำหนดน้ำหนักแห้งอย่างถูกต้องเป็นปัจจัยสำคัญสำหรับการดูแลผู้ป่วยที่ต้องบำบัดทดแทนไตระยะยาวด้วยการฟอกเลือดทางหลอดเลือด (Hemodialysis; HD) การประเมินน้ำหนักแห้งแบบดั้งเดิมด้วยการตรวจร่างกายและซักประวัติมีความแม่นยำ ในขณะที่การวัดที่ใช้หลักการวิเคราะห์องค์ประกอบของร่างกายจากความต้านทานของกระแสไฟฟ้า (Bioelectrical Impedance Analysis; BIA) ด้วยเครื่อง Body composition monitor (BCM) ทำให้ได้น้ำหนักแห้ง (BCM-DW) ที่มีความแม่นยำสอดคล้องสูงและเป็นวิธีมาตรฐานที่ดี อย่างไรก็ตามเครื่องมือมีราคาที่สูงและมีอยู่อย่างจำกัด จึงมีความพยายามที่จะพัฒนาเครื่องมือชนิดใหม่เพื่อใช้ทดแทน

วัตถุประสงค์ เพื่อพัฒนาโมเดลที่ใช้หลักการเรียนรู้ของเครื่อง (Machine learning; ML) ซึ่งเป็นโปรแกรมหนึ่งทางปัญญาประดิษฐ์เปรียบเทียบความแม่นยำสอดคล้องของการทำนายน้ำหนักแห้ง (Machine learning - Dry weight; ML-DW) ของผู้ป่วยบำบัดทดแทนไตระยะยาวด้วยการฟอกเลือดทางหลอดเลือด เทียบกับน้ำหนักแห้งที่วัดด้วยเครื่อง BCM

ระเบียบวิธีวิจัย: การพัฒนาย้อนหลังใช้ข้อมูลสืบค้นระหว่างปี 2560 ถึง 2565 จากสองสถาบันในกรุงเทพฯ ปัจจัยที่ใช้ประกอบด้วยข้อมูลทั่วไปของผู้ป่วย ข้อมูลห้องปฏิบัติการ ข้อมูลที่เกี่ยวข้องกับการฟอกเลือด และข้อมูลที่เป็นเกี่ยวข้องกับลำดับเวลาที่เปลี่ยนแปลงระหว่างการฟอกเลือด หลายโมเดลจะถูกพัฒนาโดยแบ่งข้อมูลออกเป็นข้อมูลกลุ่มฝึกฝนและกลุ่มที่ใช้ปรับค่าพารามิเตอร์ของโมเดล โมเดลสุดท้ายจะนำไปทดสอบกับกลุ่มทดสอบต่างสถาบัน ค่าที่ได้จากการทำนายคือ ML-DW จะถูกเปรียบเทียบความสอดคล้องกับ BCM-DW

ผลการศึกษา: รวบรวมข้อมูลการล้างไตได้ทั้งหมด 1,151 ครั้ง มีข้อมูลที่เกี่ยวข้องกับลำดับเวลา 56,000 ข้อมูล โมเดลสุดท้ายคือ Stacked ML model ผลความสอดคล้องของการทำนาย พบว่าML-DW ทำนายได้มากกว่า BCM เฉลี่ยผลต่างอยู่ที่ 0.78 กิโลกรัม และมีค่าขอบเขตของผลต่างอยู่ในช่วง -3.7 กิโลกรัม ถึง +2.2 กิโลกรัม

สรุป: เป็นการศึกษาแรกที่พยายามนำปัญญาประดิษฐ์เข้ามาทำนายน้ำหนักแห้งที่ได้จากการวัดด้วย BCM-DW มีการคำนึงถึงข้อมูลที่เป็นเกี่ยวข้องกับลำดับเวลาซึ่งเป็นข้อมูลหลักระหว่างการฟอกเลือด และมีการทดสอบนอกกลุ่มประชากร อย่างไรก็ตามโมเดลนี้พบว่า ข้อมูลลำดับเวลาไม่สามารถนำไปสู่การทำนาย ML-DW ที่ดี และการทำนายยังออกมาได้โดยสังเขป มีค่าขอบเขตผลต่างจากการวัดเทียบกับ BCM ที่ยังกว้าง จึงยังไม่อาจเข้ามาทดแทน BCM-DW ได้จริง อย่างไรก็ได้ทำให้เห็นถึงการนำปัญญาประดิษฐ์เข้ามาปรับใช้ในการรักษาผู้ป่วยโรคไต และเป็นโอกาสที่สามารถนำไปใช้พัฒนาโมเดลอื่นต่อไปได้ในอนาคต

สาขาวิชา	อายุรศาสตร์	ลายมือชื่อนิสิต .....
ปีการศึกษา	2565	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

6470018830 : MAJOR MEDICINE

KEYWORD: Dry weight, Hemodialysis, Body composition monitor, Artificial intelligence

Nataphut Boonvisuth : Development of an Artificial Intelligence Model for Prediction of Dry Weight in Chronic Hemodialysis Patients and Assessment of its Accuracy Compared to Standard Bioelectrical Impedance Analysis. Advisor: Prof. KHAJOHN TIRANATHANAGUL, M.D. Co-advisor: SIRA SRISWASDI, Ph.D.

Proper determination of dry weight (DW) is crucial for achieving positive outcomes in hemodialysis (HD) patients. However, the traditional clinical assessment of DW (C-DW) is often inaccurate. Recently, bioimpedance spectroscopy (BIS) analysis using a Body Composition Monitor (BCM) device has emerged as a gold standard method for determining DW (BCM-DW). Despite its accuracy, the high cost of the BCM device limits its accessibility. To overcome this challenge, the current study proposes a machine learning (ML) model, which is a part of artificial intelligence (AI), to assess DW using available clinical and laboratory parameters.

Objective: To develop an ML model for predicting DW (ML-DW) and compare it with BCM-DW.

Methods: The study consisted of a model development phase and a performance assessment phase. Retrospective data from chronic HD patients between 2017 and 2022 from two dialysis centers in Bangkok were retrieved. The parameters for this ML model included demographic, dialysis prescription, laboratory, and intradialytic time-varying data. The data utilized during the ML model development phase consisted of a training group for optimizing the parameters of the models and a validation group for determining when to stop the optimization. The final output of the model was ML-DW. The primary outcome of the study was the agreement comparison between ML-DW and BCM-DW.

Results: All 56,000 time-varying data from 1,151 HD sessions were included in the ML model. The mean BCM-DW was 58.8±11.7 kgs, while the mean predicted ML-DW from the model was 59.5±11.3 kgs. The Bland-Altman plot showed the bias estimated by the mean difference was 0.78 kg, and the limit of agreement was -3.7 to 2.2 kg.

Conclusion: This was the first study that developed a machine learning model aimed at predicting BCM-DW. Compared to other models, this one tried to explore the utilization of time-series data in the input variables. It also demonstrated external validation across different institutions. This study served as a proof-of-concept that machine learning can be a useful tool for DW prediction, but it is not yet a replacement tool for BCM. This warrants further model development that can be widely used in real practice.

Field of Study:	Medicine	Student's Signature .....
Academic Year:	2022	Advisor's Signature .....
		Co-advisor's Signature .....

ACKNOWLEDGEMENTS

The author would like to express sincere gratitude to Ms.Kanitha Tiankanon and her team at BKIH for their invaluable assistance in retrieving the necessary data for this research. The author is also deeply grateful to Dr.Sira Sriswasdi and Mr.Raksit Raksasat for providing valuable advice throughout the coding process, and to Prof.Kriang Tungsanga, Prof.Somchai Eiam-Ong, Prof.Yingyos Avihingsanon, Assoc.Prof.Pisut Katavetin and Dr. Nuttha Lumlertgul for their contributions to the writing process. Finally, the author would like to extend heartfelt thanks to his family and fiancee for their unwavering support throughout the journey.

Nataphut Boonvisuth

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Background and Rationale

Chronic kidney disease (CKD) is one of the most common non-communicable diseases in Thailand, with an estimated prevalence of 11.6 million cases, accounting for 17.5% of the Thai population. Approximately 130,000 patients develop End-stage renal disease (ESRD) and require renal replacement therapy. The incidence of patients on hemodialysis has persistently increased from 74 to 233 per million patients from 2008 to 2016[1].

The dry weight (DW) of each patient indicates the ultrafiltration volume in each hemodialysis session, and it is a crucial parameter linked to various adverse outcomes. Correct DW assessment can lead to better patient care, but targeting the proper DW can be troublesome. The conventional method of DW prediction is clinical assessment (C-DW), but this method is inaccurate. Physical examination, such as blood pressure measurement, jugular venous pulse, and peripheral edema, is an insensitive method for detecting underweight patients. This traditional clinical-based method not only varies between assessors but also comes with intra-rater variability, leading to the development of various tools to provide more objective data to assist physicians in setting DW.

The Body Composition Monitor (BCM; Fresenius Medical Care, Bad Homburg, Germany) is one of the most reliable tools for assessing dry weight. It uses an electrical principle based on measuring the resistance caused by an electric current passing through living tissues. In many institutions in Thailand, dry weight from BCM (BCM-DW) has been used as the best available gold standard. However, there are issues with its cost and generalizability, making a new tool necessary for proper dry weight assessment.

Machine learning, a subset of artificial intelligence, is a potential solution to this problem. However, there is a lack of studies that use BCM-DW as an output for

their model's learning and consider the hidden associations between time-series data and BCM-DW. With the increasing amount of hemodialysis in Thailand, there is a need for an accurate, operator-independent, emotionally-irrelevant, inexpensive, user-friendly, less time-consuming, and widely accessible tool for proper patient care.

Our purpose is to develop a model for dry weight prediction using machine learning, which would be the first to use BCM-DW as an output, known for its robustness and accuracy. For generalizability, the performance of this model will be tested with another institute. Additionally, this study would be the first to use time-series data as an input for the model's dry weight prediction.

1.2 Research question

Primary research question

Can machine learning which is a part of artificial intelligence accurately predict dry weight (ML-DW) in patient with ESRD on HD, with a high level of agreement compared to BCM (BCM-DW), which is considered the gold standard for dry weight assessment?

1.3 Objective

To develop the machine learning model and determine its agreement of DW by machine learning (ML-DW) and DW measured by BCM (BCM-DW).

1.4 Hypothesis

Machine learning which is a part of artificial intelligence can accurately predict dry weight (ML-DW) in patient with ESRD on HD, with a high level of agreement compared to BCM (BCM-DW) which is considered the gold standard.

1.5 Conceptual framework

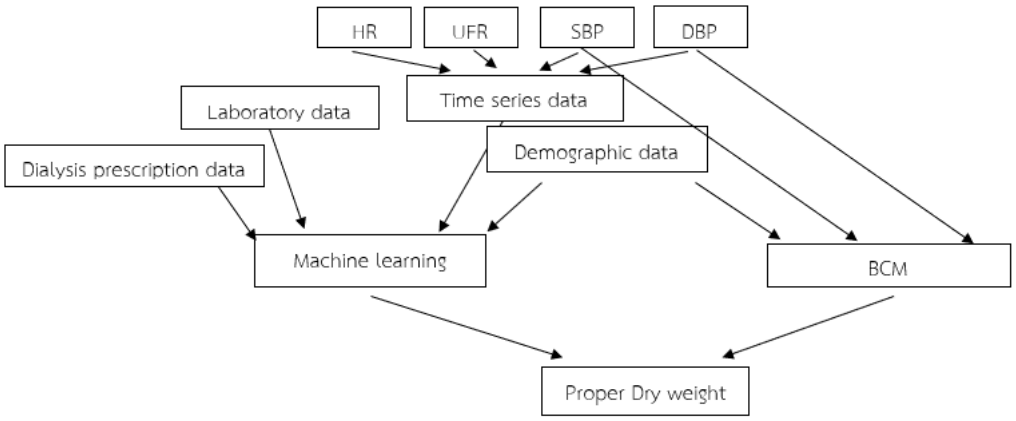


Figure 1: Conceptual framework of the study.
Abbreviations: SBP, systolic blood pressure; DBP, diastolic blood pressure; HR, heart rate; UFR, ultrafiltration rate; BCM, Body Composition Monitor.

1.6 Assumption

None

1.7 Operational definitions

A chronic hemodialysis session is defined as a dialysis session from a patient with end-stage kidney disease who has been on renal replacement therapy via hemodialysis for at least 3 months.

A stable chronic hemodialysis session is defined as a dialysis session from either prevalent or incident hemodialysis patients who have been using dialysis for at least 6 months, without infection, heart failure, cardiac arrhythmia, or sudden cardiac arrest. Patients should also meet the dialysis adequacy criteria set by KDOQI, which includes a weekly standard Kt/V of at least 2.1 or a single pool Kt/v of at least 1.2[2].

1.8 Limitation

This study is retrospective and do not include physical examination or patient symptoms, which are factors used by physicians in both institutes to adjust the patient's DW.

1.9 Expected benefit and application

The accurate results with good agreement between this model and BCM will make DW assessment more generalizable, leading to proper care of hemodialysis patients in the future.

1.10 Challenges

Since this study is utilizing machine learning for the development of the model, it requires a large amount of data, specifically dialysis session data on the same day as BCM measurement, hence, the number of available data may be limited. Additionally, developing a machine learning model requires expertise from a data scientist who can properly tune the model and set up a suitable pipeline.

1.11 Key words

Dry weight, Bioelectrical Impedance Analysis, Body composition monitor, Artificial intelligence, Neural network, Chronic Hemodialysis

CHAPTER 2

LITERATURE REVIEW

2.1 Dry weight for hemodialysis patient

DW in the context of hemodialysis refers to a patient's target weight that accounts for the residual fluid in their body after dialysis has removed excess fluid, but not to the extent of causing dehydration[3]. Knowing a patient's DW is essential for proper patient care because it is not just a number but also linked to many adverse outcomes. Overhydration status is linked to various problems, including increased blood pressure and a higher risk of cardiovascular issues, which are linked to mortality[4]. Conversely, being underhydrated can result in negative consequences, such as intradialytic hypotension, ischemic organs, increased recovery time following dialysis, and decreased residual renal function[5, 6].

2.2 Method for dry weight assessment

The conventional method of determining DW through clinical assessment (C-DW) combines many skills, such as history taking, symptom evaluation, physical examination, blood pressure monitoring, jugular venous pulse evaluation, and peripheral limb edema assessment. However, many of these parameters are subject to variability, both between different assessors and within a single assessor over time. Despite the development of clinical scores, vagueness still remains[7]. The use of blood pressure as a surrogate measure is unreliable due to volume-independent hypertension caused by age and comorbidities[8]. On the other hand, intradialytic hypotension, which may indicate underhydration, can be influenced by various factors that disrupt plasma volume refilling, such as serum sodium, calcium, hemoglobin, or albumin[9, 10]. Another promising parameter that reflects volume status is the variation of heart rate during dialysis. It has been suggested as a compensatory mechanism for the body's homeostatic balance in response to fluid status. A small study found that heart rate followed a sigmoid curve as volume status became more

depleted, but further research is needed to confirm these findings[11]. Cramping and fatigue are unreliable and not specific to underhydration, while clinical scores for determining overhydration require the accumulation of at least 2-3 liters of fluid[12]. As a result, efforts have been made to incorporate additional tools to provide more objective data to aid healthcare providers in setting an accurate DW. While many tools have been developed for DW assessments, accuracy and availability remain problematic[13]. The best accuracy method for volume assessment is Extracellular fluid and total body water measurement via Sodium bromide (NaBr) and deuterium (D2O). Nevertheless, this method has limitations due to its invasiveness and limited availability.

Assessment method	Accuracy	Other limitations
Clinical assessment	Low	Lack standardization, high variability
Chest radiography	Low (hypervolemic)	Minute radiation exposure
Serum NT-proBNP	Low (hypervolemic)	Variable cost, Not available at dialysis unit
Echocardiogram	Good	Effect by valvular heart disease, Operator dependent, Not available at dialysis unit, High cost, Time burden
Inferior vena cava diameter	Low	Operator dependent, Low repeatability, Time burden
Lung water ultrasound	Good (hypervolemic)	Operator dependent, Not available at dialysis unit, Time burden
Bioelectrical Impedance Analysis	Good	High reproducibility, Not available at dialysis unit, Costly
Relative blood-volume monitoring	Good	High reproducibility, Not available at dialysis unit, Costly

Extracellular volume (NaBr)	Good	Invasive, High cost, Time burden, Not available at dialysis unit
Total body water (D2O)	Good	Invasive, High cost, Time burden, Not available at dialysis unit

Table 1: Comparison of dry weight assessment by each method.
Abbreviations: N-terminal (NT)-pro hormone B-type natriuretic peptide, NT-proBNP; Sodium Bromide, NaBr; Deuterium, D2O; Adapted from Jennifer E Flythe, et al[13].

2.3 Body composition monitor

In 1963, Thomassett introduced a new technique called Bioelectrical Impedance Analysis (BIA)[14]. BIA uses two electrodes attached to the human body to conduct electrical current through the tissue. The low frequency alternating current causes conduction almost exclusively through the extracellular spaces of the tissues, while the high frequency range allows current to pass through both the intracellular and extracellular spaces. Thus, it can measure the resistance and reactance of both intracellular and extracellular water. Combined with the body compartment model, which includes tissue mass normohydrated lean, normohydrated adipose tissue mass, and overhydration, it can determine the patient's DW. BCM is a tool that uses the BIA method for measurement. It retrieves the electrical response of 50 different types of frequencies between 5 and 1,000 kHz. BCM assumes a division of TBW into extracellular water (ECW) and intracellular water (ICW), which are estimated using specific software provided by the manufacturer from the equation by Moissl et al[15]. Excess fluid is calculated as the difference between the measured and expected ECW in a normal situation[15, 16].

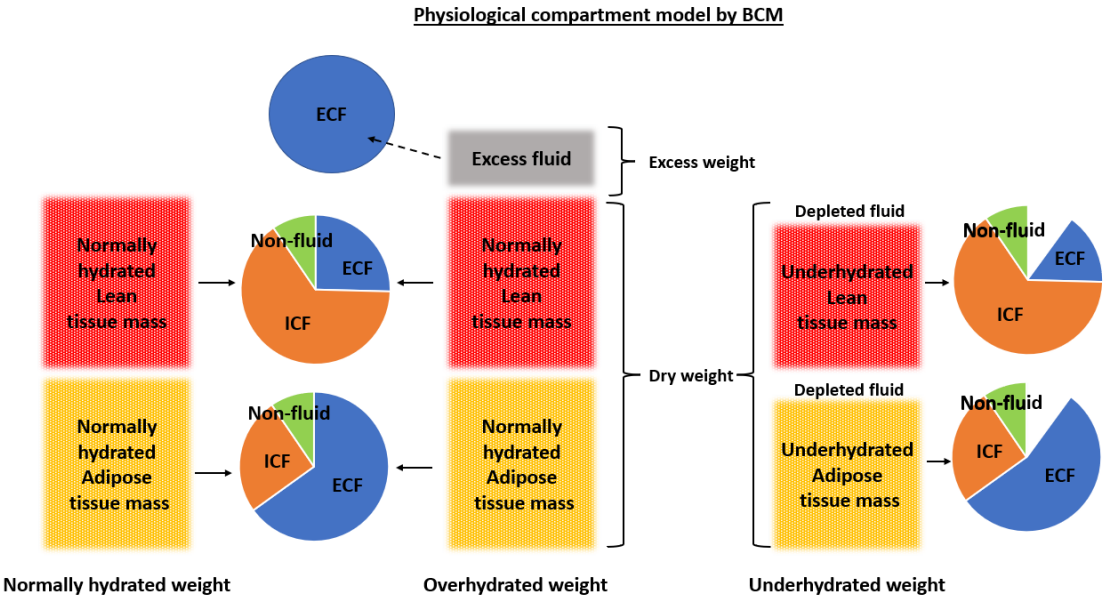


Figure 2: Distribution of water in different hydration status.
Abbreviations: BCM, Body composition monitor; ECF, extracellular fluid; ICF, intracellular fluid (adapted from Chamney, et al[16]).

- A person who has a normally hydrated weight has water in their ECF distributed in two compartments: normally hydrated lean tissue (comprising muscle, bone, and other organs) and normally hydrated adipose tissue. These two compartments will be the same as a reference population at the same age and sex, and this is called the dry weight (DW).
- A person who has an overhydrated weight has excessive ECF from the two compartments mentioned above. This excess fluid is equal to present weight subtracted by DW.
- In contrast, a person who has an underhydrated weight has ECF lower than the two compartments mentioned above. This depleted fluid is equal to DW subtracted from the present weight.

Therefore, the BCM demonstrates the DW of each patient and the excess or depleted fluid in that stage. The BCM uses bioelectrical impedance analysis (BIA) to

measure the water distribution in the body, and this can provide an objective measurement of a patient's hydration status.

2.4 Body composition monitor: Its clinical implementation and limitation

The performance of BCM has been validated by measuring extracellular fluid and total body water through Sodium bromide (NaBr) and deuterium (D2O in both healthy populations and dialysis patients[15, 17]. The tool has other strong points such as high repeatability and ease of use. Since then, it has been widely accepted as the best available gold standard, with much higher accuracy than C-DW. Several studies have demonstrated that BCM-DW can identify overhydrated patients, around 23-26% of those diagnosed as euvolemic by clinical assessment[18, 19]. Additionally, various studies have found that using BCM-DW has benefits over C-DW in several clinical outcomes, such as:

- Blood pressure

  An RCT from Romania found that using BCM-predicted DW leads to better blood pressure control in dialysis patients. After 2.5 years of follow-up, the systolic blood pressure was 2.43 mmHg lower (95% CI, −7.70 to 2.84) (P=0.40)[20]. Another RCT from China, BOCOMO, found that BCM-guided DW for 3 months can lead to a decrease in systolic blood pressure of around 6 mmHg[21]. A recent systematic review and meta-analysis from the UK, which included five randomized controlled trials in HD patients, found that bioimpedance-based DW assessment was significantly associated with an improvement in systolic blood pressure of -2.73 mmHg (-5.00 to -0.46 mmHg)[22].

- Mortality

  Studies found that overhydration identified by BCM-DW, defined by ECV>15% or approximately 2.5 kg, increased the hazard ratio of mortality by 2.1 times after 3.5 years of follow-up[23]. A cohort from Poland also found that BCM, which identified overhydrated status, was one of the significant factors

associated with higher mortality in the long-dialysis vintage (>2 years) subgroups[24]. While the use of BCM-guided DW was related to proper DW targets and a decrease in blood pressure, which had a positive effect on survival. Nevertheless, studies proving the benefit of BCM-DW on mortality were still lacking. One meta-analysis found that the use of BIA had no significant effects on mortality (HR 0.69, 95% CI 0.23 to 2.08; p = 0.51, I 2 = 54%)[25]. The latest RCT from China also found that BCM-DW showed no difference in survival by the Kaplan-Meier curve (HR = 0.51, 95% confidence interval: 0.24–1.08, log-rank test p-value = 0.07). However, the follow-up period was only 13 months. Interestingly, this curve showed an increasing trend of survival improvement in the BCM group compared to the control group over time[21].

- Cardiac parameters

  Left ventricular mass index (LVMI) is one of the factors that strongly linked to mortality[26]. An RCT found that after 12 months of applying BCM-DW, LVMI decreased from a baseline of 131 (SD 36) to 116 (SD 29) ($p < 0.001$), while C-DW remained the same at 120 (SD 20) ($p = 0.9$). However, changes in parameters such as left ventricular hypertrophy (LVH) were not significant[27].

- Intradialytic adverse events

  An RCT from Taiwan showed benefit of BCM-DW over C-DW in decrease incidence of intradialytic hypotension (6.10 vs. 6.62 %, $p < 0.05$). Probing toward BCM-DW was done at pace of 0.2–0.5 kg change per week. Additionally, there were no statistically significant differences in the incidence of other adverse events during dialysis, such as cramping, fatigue, and other patient-reported symptoms, between BCM-DW and C-DW[28].

Although BCM has been shown to accurately determine DW and provide several benefits, it is important to note that the tool has some limitations. For instance, BCM may not be suitable for patients with certain medical conditions, such as those with implanted electronic devices or amputated limbs, as it may provide

inaccurate measurements. Additionally, there are a few things that need to be mentioned.

- Body mass index (BMI)

  The study indicated that BCM measurements can cause excessive variation in the ECW compartment in individuals with extremely high or low BMI[29]. Accordingly, guidelines recommended applying this method to study populations with a BMI between 16-34 kg/m2[30]. Therefore, only patients with BMI within the acceptable range were included in this study.

- Age of population

  Since dialysis patients tend to be elderly, while the accuracy of BCM for DW determination has only been validated in populations with a limited age range, typically around 86 years old. To address this limitation, we only included patients who were 86 years old or younger in our study.

- Timing for measurement

  BIA measures the body's resistance and reactance to electricity, and there are a few things to consider regarding the timing of measurements. Firstly, water fluid in body compartments can shift between compartments, so BCM measurement after a dialysis session may be more appropriate than pre-dialysis measurements[31]. However, post-dialysis BCM measurements may have drawbacks due to interference from changes in electrolyte composition, although its clinical impact is unknown[32]. In 2004, a study found that BIA variability would be stable for at least 120 minutes after a dialysis session[33]. Later, guidelines recommended post-dialysis measurements for at least 30 minutes. If pre-dialysis measurements are to be taken, this parameter should be used as a trend by serial measurements, rather than a static value. Despite these considerations, some studies have found good agreement between body water composition in pre-dialysis measurements and extracellular water measurements by Bromide, with a mean difference of 0.8 kg[15]. In real clinical practice, it may not be feasible for dialysis centers to ask patients to stay for another two hours after a session. Therefore, all institutes in Thailand that

use BCM measure DW before dialysis, along with other clinical parameters. This retrospective study also retrieved BCM data from pre-dialysis measurements.

- Cost and Generalizability

  The major limitation of BCM is its cost, with the machine costing around 150,000 baht and each measurement requiring a specific electrode costing around 300 baht. This limits its use in Thailand mostly to university hospitals, and only nine provinces have BCM machines, with most having only one machine per province. A cost-effective analysis from the United Kingdom in 2021 found the ICER for bioimpedance-guided fluid management versus standard management to be £16,536 per QALY gained, which was acceptable for the United Kingdom's threshold for ICER capped at £20,000 per QALY[34]. However, in Thailand, the ICER threshold for health policy is 160,000 baht per QALY, so the cost of £16,536 (690,000 baht) per QALY is too high to apply this tool to all patients.

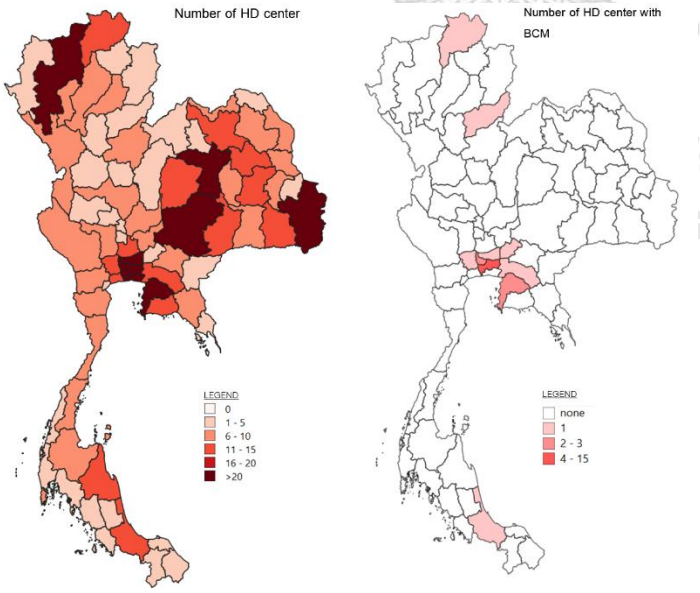


Figure 3: Demonstrated amount of hemodialysis centers (Left) and number of HD centers with BCM (Right) in Thailand.

Abbreviations: HD, hemodialysis; BCM, Body composition monitor. Adapted from "Annual Report Thailand Renal Replacement Therapy 2020" By The Nephrology Society of Thailand and detailed record of BCM's sale performance between 2018-2022.

In conclusion, there is a significant gap between the number of good tools and the demands for DW assessment in Thailand. Therefore, a new tool that concern about these limitations while preserve the same performance as BCM is needed to be developed for good care of dialysis patients.

2.5 Machine learning model and dry weight prediction

Machine learning (ML) is a subset of artificial intelligence that allows computers to learn from data and improve without explicit programming. These algorithms process input and output data to develop models[35]. By the way, deep learning, a type of machine learning, uses multiple layers of non-linear modules to create more complex functions. This substructure is beyond human ability, leading to previously unimaginable solutions for complex tasks and the discovery of hidden associations[36]. These cutting-edge technologies result in an assistive tool for physicians to make decisions with more cost-saving and less time-consuming.

Few studies have used machine learning for the prediction of dry weight (ML-DW). However, most of them compare their results to clinical assessments (C-DW), which are subjective and prone to errors. Additionally, some studies have used unacceptable rationale. For instance, a study by Olivier Niel et al. used BCM-DW as an input variable, which is considered the gold standard for predicting DW, and then predicted the result of C-DW. The study showed good agreement, but it did not bring any benefit to clinical implementation[37]. Another study by Hae Ri Kim et al. used BCM-DW as an input variable and predicted the final result in the form of an absolute gap between BCM-DW and C-DW. As a result, the effect of the input variable of BCM was again contaminated in the prediction process. The interpretation

was also difficult to do, since absolute gap did not tell the direction whether it was overhydrated or underhydrated[38].

Furthermore, comparing the performance of different machine learning models that predict the same thing can be challenging. The proper way to compare a new model with an existing one is to see if they agree with each other. However, some studies used tests like accuracy[38] or correlation[39], which cannot determine if the models are truly similar due to their inability to account for bias or the range of possible results. Such tests are not good for comparing with a "gold standard" test[40].

Few studies have used correct statistical methods. However, it compared machine learning-based dry weight prediction (ML-DW) with clinical dry weight assessments (C-DW). Olivier Niel et al. found that their ML model was able to predict C-DW with good agreement, with a mean difference of 0.09% of C-DW. Although, there was a wide limit of agreement at -4.27 to +4.44% of C-DW (absolute kilogram was not reported), and the number out of agreement was 20 out of 476 samples[41]. The input parameters in this study included demographic data, and time-variable data as a time stamp of pre-dialysis blood pressure and heart rate was used as an input.

In another study by the same author, the same input and output were used with a different ML model. The mean difference was nearly the same (-0.04% of C-DW), and the number out of agreement was lower at 17 out of 476 samples. However, the limit of agreement was still wide (-4.41 to +4.33% of BW, absolute kilogram not reported)[42].

In conclusion, none of these studies used ML to predict the more objective and reliable method of DW assessment, such as BCM-DW. Additionally, none of them used time-series variables, which are the most abundant data in each dialysis session for prediction. As mentioned earlier, intradialytic changes of vital signs may have a complex relationship with the patient's volume status, so these variables may be an

essential factor for DW prediction. Moreover, the agreement test results sometimes did not provide an absolute weight number, which could be difficult to use in real clinical implementation. Lastly, none of these models were externally validated with data outside their institute. Table 2 provides the details of each study.

Study	Input	output	Patient	model	Result
Hae Ri Kim, et al. 2021	Demographic Dialysis Laboratory Time variable (Time stamp) BCM-DW C-DW	Absolute difference between BCM-DW and C-DW	1672	XGBoost	Accuracy 28.54 % Mean absolute error 1.29 kg
Jainn-Shiun Chiu, et al. 2005	Demographic	BIA-Total body water (not DW)	54	ANN	Pearson's correlation coefficient (-0.911) RMSE 2.48 kg
Olivier Niel, et al. 2018	Demographic Dialysis Time variable (Time stamp) BCM-DW	C-DW	14	Multilayer perception neural network	Mean difference +0.497 kg (LOA -1.33 to +1.29 kg)
Xiaoyi Guo, et al. 2021	Demographic Time variable (Time stamp)	C-DW	476	Sparse Laplacian regularized Random Vector Functional Link	Mean difference 0.09% of C-DW, (LOA -4.27 to +4.44 %) Number out of agreement: 4.25%
Xiaoyi Guo, et al. 2021	Demographic Time variable(time stamp)	C-DW	476	Multiple Laplacian-regularized radial basis function networks	Mean difference: -0.04% of C-DW (LOA -4.4 to +4.33%) Number out of agreement 3.57%

Table 2: Descriptive data of available trial about machine learning model and dry weight prediction.

Abbreviations: BCM-DW, Body composition monitor adjusted dry weight; C-DW, Clinical adjustment dry weight; BCM, Body composition monitor; XGBoost, Extreme Gradient Boosting; ANN, artificial neural network; RMSE, Root mean squared error; LOA, limit of agreement

2.6 Machine learning model for time-series data and its structure

Dealing with time-series data can be troublesome. Conventional statistical models such as ARIMA (Autoregressive Integrated Moving Average) or SARIMA (Seasonal Autoregressive Integrated Moving Average) are commonly used for time-series data, but these are primarily used for forecasting or predicting. They are designed to model the underlying patterns in time-series data and make predictions of the same parameter in the future based on those patterns. Thus, they are not suited for classification tasks or prediction of value of different parameter. To approach this problem, a recurrent neural network is needed.

Recurrent Neural Networks (RNNs) are a type of neural network designed for processing sequential data, such as time-series data or natural language text. The key characteristic of RNNs is their ability to retain information from past inputs, which allows them to model temporal dependencies and contextual information in sequential data[43]. The basic structure of an RNN involves a recurrent hidden layer that is connected to itself through time. The inputs are processed in parallel at each time step and are fed through the hidden layer, which generates an output and updates its internal state. The internal state of the RNN can be thought of as a memory that allows it to capture information from past inputs and use that information to generate the current output. The output of this could be either a prediction or a classification. This makes RNNs well-suited for a wide range of tasks, including language modeling, graph prediction, and speech recognition. There are several variants of RNNs, such as GRU (Gated Recurrent Unit) and LSTM (Long Short-Term Memory).

GRUs have been designed to be a simple and fast type of RNN, with fewer parameters to train, which makes them more efficient for some tasks. However, there are drawbacks such as poor preservation of long-term dependencies in sequential data, which can cause the model to forget these dependencies over time or result in vanishing gradients from backpropagation of errors through time[43].

LSTM is an improved RNN architecture that can construct a long-term sequence relationship and has dominant memory units, which can solve the problem of gradient disappearance or explosion in GRU[44]. Thus, it will be chosen as one of the core models for this study.

Structure of LSTM

The memory cells in LSTM enable the transmission of gradient information over long distances during training. In addition, three gates (input gate, forget gate, and output gate) are established. The detailed structure of LSTM is shown in figure 4.

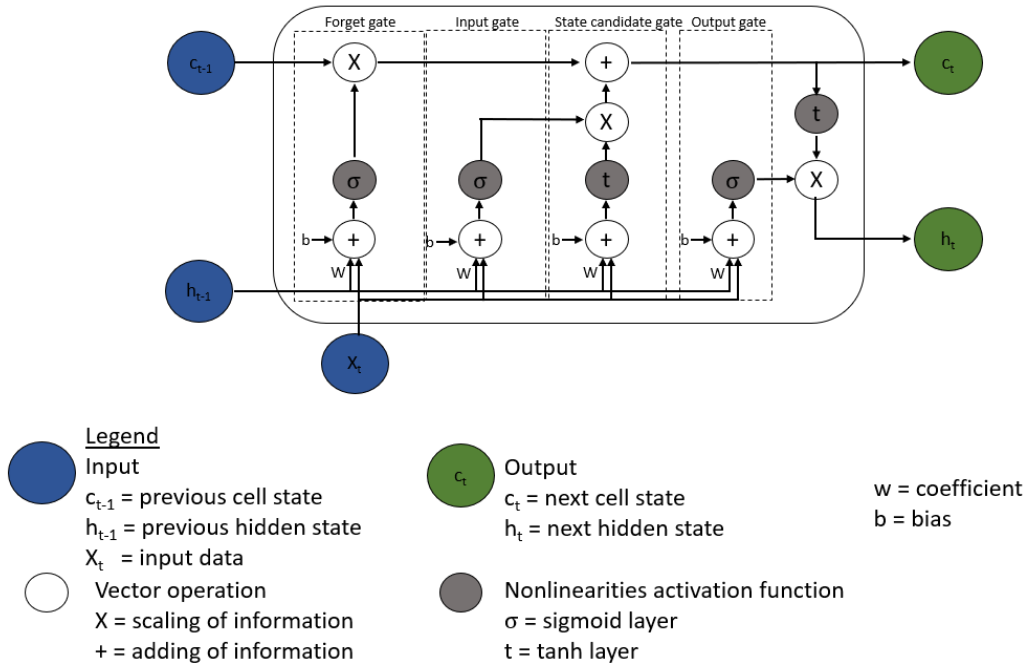


Figure 4: Structure of LSTM model. Adapted from Hochreiter, et al[44].

The forget gate, presented in Equation 1, decides the retention or discard of information. The sigmoid function is applied to the previous hidden state and the current input simultaneously. The output value, f_t, is between 0 and 1, with values close to 1 retained and values close to 0 discarded.

$$f_t = \sigma(W_f [h_{t-1}, X_t] + b_f)$$

Equation 1: f_t is the forget gate output. X_t and h_{t-1} are the previous hidden state and current input, respectively. W_f and b_f are the coefficient and bias values for the forget gate, respectively.

The input gate updates the cell state, determining how many network inputs are retained in the current cell state. The function of the input gate is shown in Equation 2. The previous hidden state and the current input are the input data transferred to the sigmoid function for adjusting the values. The output value zero means unimportant, while one means essential. In addition, the tanh function is used for the previous hidden state and the current input to generate a candidate vector, c'_t, as shown in Equation 3. Finally, the cell state, c_t, presented in Equation 4, is the output value from the sigmoid and candidate vectors.

$$i_t = \sigma(W_i [h_{t-1}, X_t] + b_i)$$

$$c'_t = \tanh(W_c [h_{t-1}, X_t] + b_c)$$

$$c_t = (f_t c_{t-1}) + (i_t c'_t)$$

Equation 2-4: i_t is the input gate output. c'_t and c_t are candidate vector and cell state values, respectively. Wi and bi are coefficient and offset values for the input gate. Wc and bc are coefficient and offset values for the candidate vector.

The output gate, as shown in Equation 5, determines the next hidden state value. The previous hidden state is passed to the sigmoid function together with the current input, and the new cell state is passed to the tanh function. The current hidden state is the product of the output of the tanh function and output gate value, as shown in Equation 6. The new cell state and current hidden state are transferred to the next step.

$$o_t = \sigma(W_o [h_{t-1}, X_t] + b_o)$$

$$h_t = o_t \tanh (c_t)$$

Equation 5-6 o_t is the output gate output. h_t is the current hidden state value. w_o and b_o are the coefficient and bias values for the output gate, respectively.

LSTM is commonly used for regression analysis. However, it is also widely used for classification tasks. This is done by adding a dense layer on top of the LSTM layers, which takes the output of the last LSTM layer and maps it to the desired number of output classes. For instance, final layer of the LSTM for regression model uses linear activation function, which produce continuous numerical output value, such as a temperature or stock price. Meanwhile, LSTM for classification tasks commonly uses sigmoid activation function for binary classification. If it is a single label, multiple classes classification, the cross-entropy loss function is selected.

LSTM has become a popular choice for sequence modeling tasks due to its ability to handle long-term dependencies and its ability to learn and remember patterns from the input data. It has been successfully applied in various fields such as speech recognition, natural language processing, and time-series prediction. Additionally, LSTM models can be adapted to handle different types of input data, including images and graphs. LSTM can also be combined with other types of neural networks, or attention-based models, to improve performance on specific tasks. Overall, LSTM has proven to be a powerful and versatile tool for a wide range of applications.

2.7 Machine learning model and neural network model for prediction and its structure

The conventional method for dealing with prediction model between many dependent variables and one independent variable can be basically done by linear regression. It is a good starting point for many problems and works well when the relationship is linear, and the model's complexity is not too high. However, for more complex relationships, several other ML models are employed for regression

analysis, such as Least Absolute Shrinkage and Selection Operator (Lasso)[45], Elastic Net (ENet)[46], Kernel Ridge Regression (KRR)[47], Extreme Gradient Boosting (XG Boost)[48] and Light Gradient Boosting Machine (Light GBM)[49].

Lasso is a modification of linear regression model that can be used for feature selection and regularization to prevent overfitting. It has better strengths in feature selection and handling high-dimensional data. It can also help with overfitting by adding a penalty term. However, it is sensitive to outliers and can be unstable when there is high collinearity between features.

ENet is another modification of linear regression model that combines the Lasso and ridge regression techniques. It can handle the issue of multicollinearity between features but requires higher computation and may not be suitable for large datasets.

KRR is a kernel-based linear regression method that can capture nonlinear patterns in the data. It can handle both linear and nonlinear relationships. Its weakness depends on its sensitivity to the choice of kernel function.

XGBoost is tree-based ensemble models that can handle nonlinear relationships and interactions between features and can improve performance through boosting techniques. It can handle missing values and reduce overfitting very well. However, it can be computationally expensive and may require careful tuning of hyperparameters.

LightGBM is another tree-based ensemble models. Its fast and efficient gradient boosting method that can handle large datasets. Contrary, it may not perform well with small datasets and may require careful tuning of hyperparameters.

Stacked regression, also known as stacked generalization, is an ensemble method that combines the predictions of several models to improve the performance[50]. It involves training several base models on the same training data and then using their predictions as input features for a higher-level model, which makes

the final predictions. Stacked regression can be used to combine different types of models and can be a good choice when the relationship between the variables is more complex or when the base models have different strengths and weaknesses. However, it requires more data, more computing resources, and more time to train.

Neural network models are also commonly used for prediction tasks. These models are built using multiple layers of interconnected nodes, with each node performing a mathematical operation on its inputs and passing the result to the next layer. The nodes in each layer are typically connected to all nodes in the previous and next layers, creating a dense network of connections. The output of the final layer is the predicted value. Neural networks can handle complex relationships between variables and can capture nonlinear patterns in the data. However, they may require more data and more computation resources to train than other models. Additionally, choosing the appropriate architecture and hyperparameters for the network can be challenging.

2.8 Rationale for choosing institute for study population

This study needs data of dialysis session together with data of BCM measurement. Collection of time-series data from dialysis sessions is a laborious task. Since each dialysis session will contain around 50-60 numbers of interdialytic time-series variable which usually be recorded by HD nurse manually. Retrieving these variables manually is both laborious and prone to human error. Therefore, to collect the required data, we needed an institute that records these parameters via electronic methods, such as the Therapy Support Suite (TSS) from Fresenius Co. We also required an institute that provides BCM measurement for labeling the outcome of interest, which few hospitals had this. Only King Chulalongkorn Memorial Hospital (KCMH) and Bhumirajanagarindra Kidney Institute Hospital (BKIH) met these criteria, but they differ in several ways that may affect the results and analysis.

KCMH is a tertiary-care and university hospital that typically handles complicated cases, and currently has around 50 active patients. As BCM measurements are performed more frequently than in general hospitals, the number of dialysis sessions with BCM measurement is high despite the relatively small number of patients.

BKIH is a private hospital that specializes in hemodialysis care and has over 500 active patients. However, BCM measurement is costly, resulting in only 2-3 measurements being performed in each patient. Therefore, the number of dialysis sessions with BCM measurement is high due to the large number of patients.

The aim of this study is to develop a new model and evaluate its performance through an external test. Additionally, a comparison between the characteristics of two institutes, with similar numbers of dialysis sessions but different patient populations, will be conducted. The study will test whether a model trained on the larger patient population can accurately predict outcomes for the smaller patient population, and vice versa. This analysis will serve as a robustness test of the model before generalization to real-world practice.

CHAPTER 3

METHOD

3.1 Research design

Diagnostic Retrospective study

3.2 Research methodology

3.2.1 Target population

This study focused on the number of dialysis sessions, as we use each session for the model's learning separately. Thus, it requires dialysis sessions with BCM measurement on the same day from ESRD patients.

Study population

Dialysis session from ESRD patients at two institutions, King Chulalongkorn Memorial Hospital (KCMH) from January 1, 2017, to December 31, 2021, and Bhumirajanagarindra Kidney Institute Hospital (BKIH) from January 1, 2018, to December 31, 2022.

3.2.2 Inclusion criteria

(a) Dialysis sessions from patients aged 18-86 years

(b) Dialysis sessions from patients who had done 2-3 sessions per week

(c) Stable dialysis, defined as prevalence cases or incidence cases with a duration of dialysis for more than 6 months

(d) BCM measurement was done on the same date as the dialysis session.

Exclusion criteria

patients were excluded if they had:

(a) Recently active and uncorrected heart disease (ischemic heart disease within 6 months, heart failure within 6 months, moderate to severe valvular disease)

(b) Atrial fibrillation

(c) An amputated limb

(d) Cirrhosis

(e) Been pregnant or breastfeeding

(f) A BMI <16 kg/m2 or >34 kg/m2.

For those eligible patients, we also exclude dialysis sessions that had:

(g) Intradialytic adjustment of dialysate concentration, including dialysate profiling or

(h) Corrupted data without a complete record.

3.3 Sample size calculation

$$n = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{[(P_0)\ (1-P_0)\beta^2]}$$

type I error (α) = 0.05

type II error (β) = 0.20

P_0 = 0.9 from Moissl, et al[15]

Dropout rate 10%

N = 130

The conventional method for predicting dry weight (DW) using logistic regression, as described by Hsieh et al., found that the new tool can be comparable to BCM in a study of 130 patients. However, in this study, a machine learning model was proposed for prediction and each dialysis session was utilized separately for model learning. The focus was on the number of dialysis sessions, as a large amount of data was crucial for the learning process. The expected amount of 1000 dialysis sessions was deemed sufficient for ML learning to answer the clinical question. Although a neural network model may have benefited from a higher number of dialysis sessions, it was acknowledged that obtaining more than 1000 BCM

measurements may be unrealistic and inconvenient. Nonetheless, the potential of a neural network was explored and its performance was assessed while keeping this limitation in mind. The team appreciated the understanding and did their best to ensure that the approach yielded meaningful results.

3.4 Data collection

The data collected for this study includes demographic data (age, sex, height, pre-dialysis weight, post-dialysis weight, and underlying diseases (hypertension, diabetes, stroke, chronic lung disease, and gout), laboratory data (hemoglobin, sodium, calcium, and albumin), dialysis prescription data (composition of sodium, potassium, calcium, bicarbonate, and temperature), time-series data during dialysis sessions (including timestamps for systolic and diastolic blood pressure, heart rate, and ultrafiltration rate), and DW data from BCM.

Demographic data were collected based on physician reports and ICD-10 codes, while laboratory data for each dialysis session were selected based on the closest sampling prior to that session, but not more than one month prior. Dialysis prescription and time-series data were collected from the TSS. In both institutes, timestamp data were monitored before the start of dialysis, for the next 30 minutes, and then every hour. However, if vital signs were unstable or the physician required, monitoring was done more frequently. BCM data in this study was pre-dialysis measurement, collected from the TSS record. Quality of measurement and Cole-Cole plot of BCM were not available for this retrospective review. After measurement was done, physicians would prescribe the target weight for that dialysis session based on a combination of both BCM-guided data and clinical assessment. So Post-dialysis BW does not equal to BCM-DW. Probing of weight was done in order to avoid rapidly change in weight that may cause consequences. Adverse events during a dialysis session, were not reported in the record and not present in this study. Intervention

required for a patient during a dialysis session was also not included as a parameter. The diagram of data retrieval is shown in Figure 5.

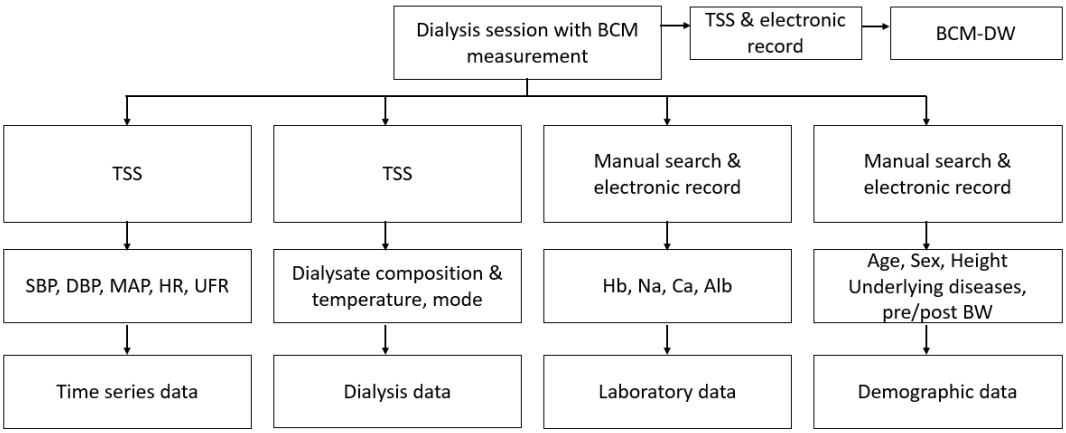


Figure 5: Flow for retrieving data.

Abbreviations: BCM, Body composition monitor; TSS, Therapy support suite; BCM-DW, dry weight from BCM; SBP, systolic blood pressure; DBP, diastolic blood pressure; MAP, mean arterial pressure; HR, heart rate; UFR, ultrafiltration rate; Hb, hemoglobin; Na, sodium; Ca, calcium; Alb, albumin; pre/postBW, pre-dialysis and post-dialysis body weight

3.5 Data analysis

Data from four domains, including demographic, laboratory, dialysis prescription, and BCM data, were combined into a single sheet. While data from time-series data was stored in another sheet.

3.5.1 Data preprocessing

The raw data of the dialysis sessions were cleaned by removing null values. As different magnitudes of data can cause larger variables to dominate over smaller ones, all ratio and interval data such as height, weight, and dialysate sodium were normalized by the StandardScaler command in TensorFlow. All features were

transformed to fall within the range of 0 (lower bound) and 1 (upper bound). Parameters that were not normally distributed were transformed using log N or Box-Cox transform. Categorical data were marked and encoded as numerical labels using LabelEncoder from the 'sklearn.preprocessing' module, and then encoded into a set of binary variables (0 or 1) for each category using the 'get_dummies' function from the pandas module. This helped the regression model deal with categorical variables. Time-series data was scaled in the same way as above. Due to the unequal time frame between each dialysis session, zero padding was used prior to each time frame to create a matrix of the same size without affecting the model's performance [51]. The output was labeled for BCM-derived DW.

3.5.2 Training, Validation and Testing set

All data from dialysis sessions were randomized and split using a command in Tensorflow into a training set (80%) to optimize the parameters of artificial neural network models, a validation set (20%) to determine when to stop the optimization process. During the splitting process, we ensured that dialysis sessions from the same patient were placed in the same set to prevent the learning model from being exposed to input, such as demographic data, that would be the same for patients who had multiple dialysis sessions and BCM measurements. A testing set was used to evaluate the performance of the final model. As mentioned earlier, we conducted two external validation configurations for testing the robustness of our model.

- Configuration A, we used the KCMH database as the training and validation set and applied the model on the BKIH database for testing.
- Configuration B, we used the BKIH patient database to train our model and then applied our algorithms on the KCMH database.

The flowchart of the study was shown in Figure 6, and it should be noted that the train, validate, and test sets were altered based on each configuration.

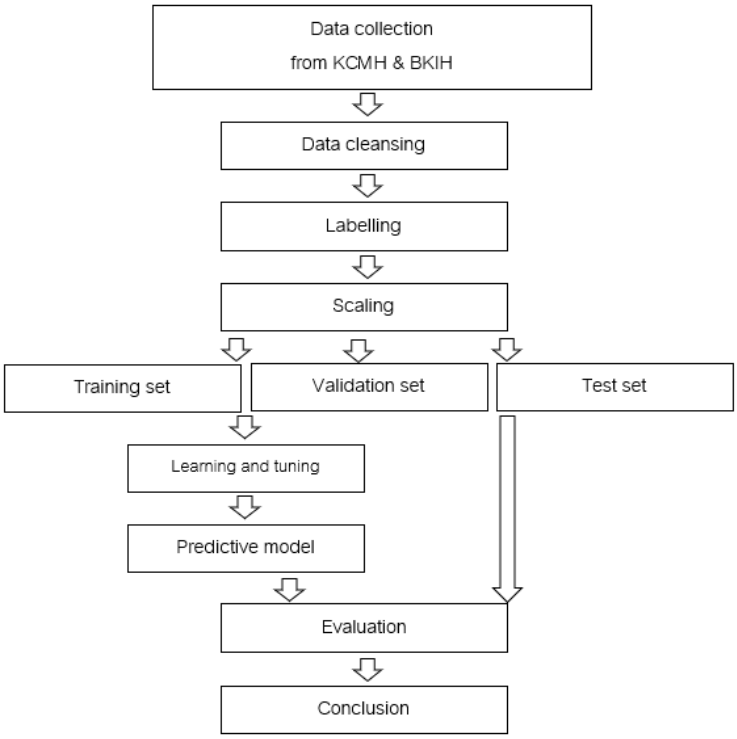


Figure 6: Flow chart of this study

Abbreviations: KCMH, King Chulalongkorn Memorial Hospital; BKIH, Bhumirajanagarindra Kidney Institute Hospital

3.5.3 Statistics and Model development phase

Descriptive analysis of input data was conducted. Data with a normal distribution was presented as mean ± standard deviation (SD). For skewed distributions, data was presented as median ± interquartile range (IQR). Associations with p values less than 0.05 were considered statistically significant. Time-series data was displayed in a descriptive manner and scatter plots were used to provide an overview of the relationship between parameters. Additionally, a histogram of BCM data was also included.

Our algorithm was built by Python v3.6.9 (https://www.python.org/) with Tensorflow backend v1.15.4 (https://www.tensorflow.org/), an open-source machine learning library. We deployed our application with this deep learning library based on

the NVIDIA container image of TensorFlow, Release 20.11 (Nvidia Corporate, Santa Clara, CA, USA) as a virtual environment.

In the model development phase, we needed to build a model that best suits our problem. Due to the difference in the type of data, the idea for model selection will be different. For example, general data (including demographic data, laboratory data, and dialysis data) contain both categorical and numerical data and can be managed by diverse models. On the other hand, time-series data requires a special model, which we have applied for LSTM. Initially, we will try to build a model by utilizing both types of data for the most efficient use. Later on, if one type of data was unsuitable for this task, we opted for only one type of data for our final model. The scheme for model selection based on the type of data is shown in Figure 7.

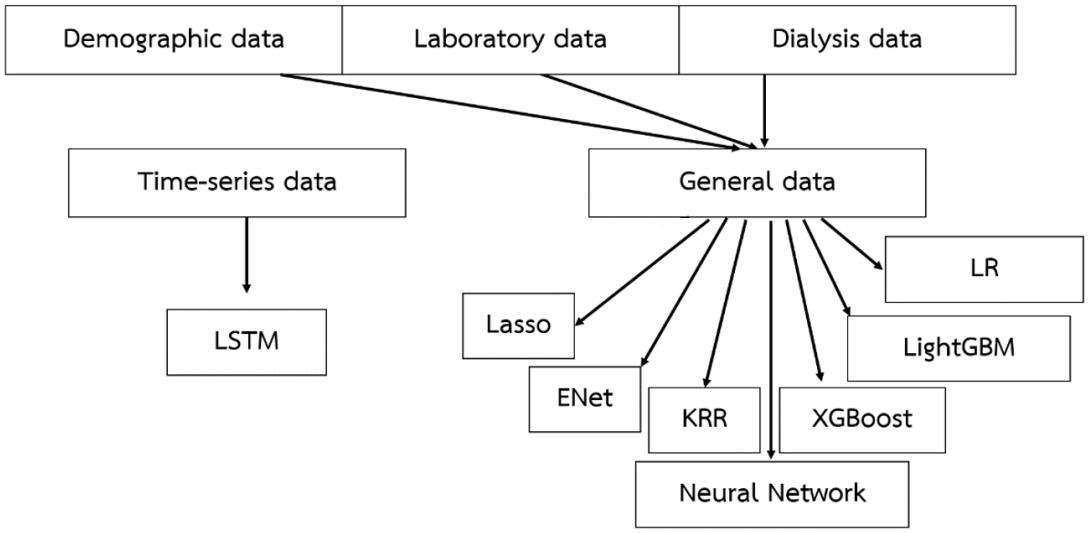


Figure 7: Type of base model for different type of data
Abbreviations: LSTM, Long Short-Term Memory; ENet, Elastic net; KRR, Kernel Ridge Regression; XGBoost, Extreme Gradient Boosting; LightGBM, Light Gradient Boosting Machine; LR, Linear regression

3.5.3.1 Processing for time-series data

As mentioned above, the bidirectional Long Short-Term Memory (LSTM) model was designed as a recurrent neural network (RNN) to capture dynamic variations in time-series data. It had capability to handle a diverse range of problems. The aim of this model is to utilize changes in vital signs and ultrafiltration rate to achieve one of the following goals:

- Directly predict or Regress BCM-DW.
- Classification dialysis sessions into different groups based on alteration of time-series data toward differences in net ultrafiltration volume.
- Reduction of high-dimensional data and transform time-series data into an embedding layer then combining with result from non-time-series data.

To develop the most efficient model, we used all of the three approaches and select the best one, if possible. The LSTM class was defined as a subclass of PyTorch's nn.Module. This class defined the architecture of the LSTM-based model that would be trained. The number of neurons in LSTM layers was varied by approach between 16-128. The learning rate was set at 0.0001. To mitigate overfitting in small dataset, drop-out at 0.1 and L2 regularization method was used with a value set to 0.001. The adaptive moment estimation (ADAM) algorithm was used for optimization with the number of iterations set to 300. The training and validation datasets were loaded into PyTorch DataLoader objects, which allowed the data to be easily fed in batches to the model during training. At the end of training, the model with the highest validation accuracy was saved.

In the first approach, called the "LSTM regression model (LSTMreg)," we trained the LSTM model to output as a predictive model. We evaluated the results in terms of mean squared error (MSE) and agreement. A smaller MSE indicates better model performance, while agreement would be shown by Bland-Altman plot. It was a graphical representation of the differences between the two methods, with 1.96 SD set as the upper and lower boundaries of the limit of agreement. A smaller mean

difference and a narrow limit of agreement indicate a good model. If the results were promising, we planned to combine this with any regression model for non-time-series data.

In the second approach, “LSTM classification model (LSTMclass)”, the LSTM model was used to classify patients into three groups based on the net UF status. Since it was known that excess UF was important factor that led to a sudden change in vital signs during dialysis, the purpose of this approach was to group dialysis sessions by the amount of net UF before sending data from each session into the specific regression model that suited each group, thereby maximizing the prediction value. Initially, the LSTMclass was responsible for screening dialysis sessions and assigning them to the appropriate groups. Since the output was a classification, the results were evaluated in terms of accuracy, F1 score (the harmonic mean of sensitivity and positive predictive value), and the confusion matrix. The confusion matrix displayed the percentage of dialysis sessions predicted in each group, with color tones indicating the accuracy of the prediction.

In the third approach, “combined time-series data and non-time-series/general data in neural network model (gtNN)”, Time-series data could be used by passing the last layer of the LSTM or the embedding layer, which provided less-dimensional data, and concatenating it with the embedding layer from non-time-series data (demographic, laboratory, and dialysis data). All of the information would pass through same optimization and dense layer. Although neural networks worked well with large amounts of data, using this approach for this small dataset may have uncovered some additional insights or trends. The result of this was continuous data of predicted weight, which was evaluated in terms of MSE and agreement.

The schematic representation of the pipeline for each approach was illustrated in Figure 8.

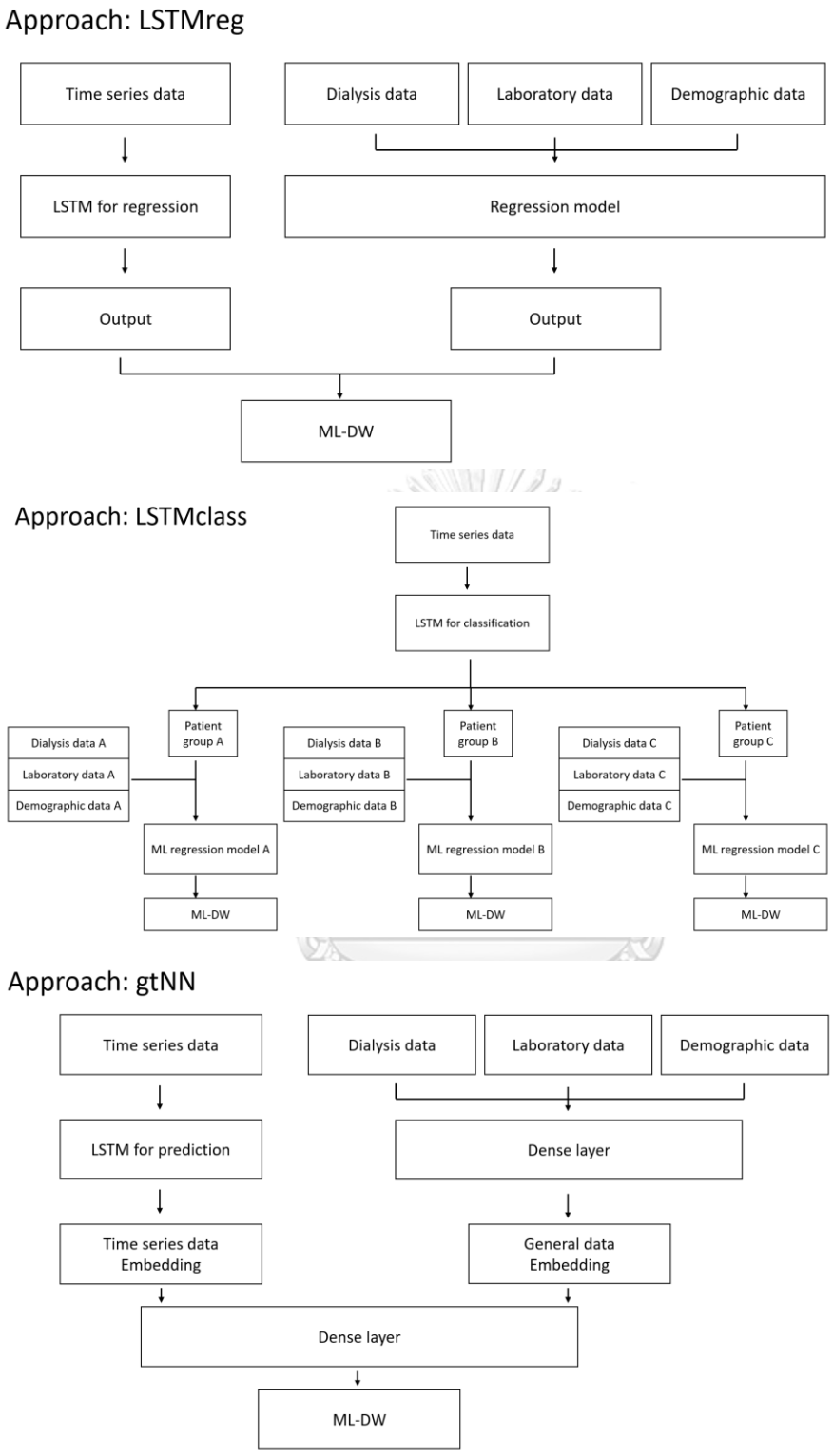


Figure 8: Different approach related to time-series data

Abbreviations: LSTMreg, LSTM regression model; LSTMclass, LSTM classification model; gtNN, combined time-series data and non-time-series/general data in neural network model; ML-DW, machine learning dry weight

3.5.3.2 Processing for non-time-series data or general data

The process for modeling of non-time-series data which aim for regression analysis was performed by three methods: conventional multiple linear regression, machine learning with stacked regression models, and neural network models. The input for these models was a combination of parameters from demographic data, laboratory data, and dialysate profile data.

For multiple linear regression (LR), the LinearRegression class in the scikit-learn library was used, and MSE was used to evaluate how far the predictions were from the actual values.

Several machine learning regression models were evaluated, including Lasso, ENet, KRR, XGBoost, Light GBM, were evaluated. Finally, a stacked regression model (STACK) was created using these base models. As the dataset was small, hyperparameters for each model were fine-tuned using the GridSearchCV() function.

- The Lasso model was created using the scikit-learn library, and normalization was performed using RobustScaler(). The strength of regularization, or alpha value, was determined using the GridSearchCV() function from scikit-learn, with values ranging from 0.001 to 0.1.
- ENet model: RobustScaler() was also applied to the data for normalization, and the alpha value was fine-tuned between 0.001 and 0.1 using the GridSearchCV() function. The l1_ratio, which determines the balance between L1 and L2 regularization, was set to range from 0.1 to 1.0.
- KRR model: This model employed a polynomial kernel function with a degree of 2, resulting in the input features being transformed to a space where each

feature is squared and multiplied by all possible pairs of features. The alpha value was determined using the GridSearchCV() function between 0.01 and 100. The constant term in the polynomial function, known as coef0, was also set using the GridSearchCV() function between 1 and 8.

- XGBoost model: To prevent overfitting due to the small datasets, hyperparameters in the XGBoost model were selected carefully. GridSearchCV() was used to select hyperparameters such as of, learning_rate (at a lower value between 0.01-0.1), max_depth (set at 3-7 to avoid overfitting), reg_alpha, and reg_lambda (both set at 0.1 to balance bias and variance). Additionally, a fraction of observations and features were randomly sampled for each tree, with values set at 0.3-0.8
- Light GBM model: To further prevent overfitting in the Light GBM model, the following hyperparameters were tuned: The learning_rate between 0.01-0.1, num_leaves between 4 to 32, n_estimators between 300-1000, min_samples_split between 10-50, max_features between 0.3-0.5, feature_fraction between 0.3-0.9, min_data_in_leaf between 10-20, and min_sum_hessian_in_leaf between 0.01 to 1.

The Group shuffle K-Fold technique was utilized for cross-validation, in which the data was partitioned into 5 parts, and the model was trained and tested on each partition. The data was randomized prior to partitioning, and a fixed random state was employed to ensure test repeatability. Each evaluation metric included MSE and SD, with lower values indicating a superior model. Finally, stacking was performed from a diverse range of base models, as their strengths and weaknesses may complement each other.

The neural network model (gNN), which used TensorFlow and Keras to predict BCM-DW based on a set of input features of non-time-series or general data, was developed. The model architecture was defined using the Keras API from TensorFlow and consisted of three hidden layers, each with 512 nodes. These layers

used different activation functions, namely sigmoid, relu, and relu. The final layer produced a single output value. The optimizer used was Adam, and the learning rate was set to 0.00001. The loss function used was mean squared error (MSE), and the model was trained for 250 epochs with a batch size of 32. The model was evaluated using MSE on the validation set.

Please note that the success of each approach in this project cannot be guaranteed, despite a rationale-driven process. Therefore, all models, including those combining time-series and non-time-series data (gtNN), those using solely non-time-series data (LR, STACK, and gNN), and those using only time-series data (LSTMreg and LSTMclass), were compared. The robustness of each model was evaluated using both configuration A and B. Mean squared error (MSE) was calculated for each model, and a Bland-Altman plot was used to compare the agreement of predictions to BCM-DW. To visualize and compare the results of each model, a Folded Empirical Distribution Function Curve (Folded EDFC) was plotted. The x-axis of the graph represents the mean difference between the predicted DW and the BCM-DW, while the y-axis defines the probability of samples with those values. The center of the graph features a vertical line that represents the reference line of zero difference, and the distance from this line and the peak indicates the estimated bias of each tool. The shape of the graph is typically mountain-like, with the base representing the entire possible range of difference between the new tool's prediction and BCM, which is linked to the limit of agreement. A steep rise in the graph indicates a higher probability of the folded variable being small, while a small hill indicates a more spread-out distribution[52]. Ultimately, the final model was chosen based on various factors including the smallest MSE, lowest mean difference, and smallest limit of agreement.

3.6 Model evaluation phase

The present study evaluated the performance of ML-DW, a continuous variable, against the established gold standard BCM-DW. Further analysis of Bland-Altman plot was conducted, as the number of samples that fell out of the limit of agreement were presented. The importance of feature was reviewed using SHapley Additive exPlanations (SHAP) values, which demonstrate the contribution of each feature to the predicted outcome. In case there were outliers beyond the limit of agreement, a sub-analysis of the data was conducted by:

- Comparing SHAP values, a summary plot of the SHAP values for samples within the range and for outliers was created.
- We reviewed the baseline characteristics of each dialysis session to determine if raw data could have led to outliers in the results.

After excluding some variables, the model was be run again to demonstrate the Bland-Altman plot. The computational efficiency of the model was also evaluated by reporting the time taken (in minutes) to execute the algorithm. The figures were generated using the Matplotlib v3.1.2[15] and Seaborn v0.11.1 Python libraries, which are renowned for their visualization capabilities in data analysis.

The full code is available at https://drive.google.com/drive/folders/1_hKlGJ9IpjtOOQii58zhV8vT_c4mZ7QV?usp=sh.

3.7 Research ethics

The protocol was approved by the institutional review board of King Chulalongkorn Memorial Hospital and adhered with the principles of the Declaration of Helsinki. It used retrospective data from medical record, hence it will not be any direct contact toward patients. All of the personal data will be stored in code that cannot be traced back. The presentation of research results will be general information. No individual information is presented.

The risk of this research is no greater than the minimal risk.

3.8 Administration and time schedule

Activities	2021				2022												2023			
	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4
1.Preparation																				
2.Data collection																				
3.Data analysis																				
4.Conclusion																				

Table 3: Time schedule of this study

3.9 Budget

List	Budget
Revenue data scientist	50,000
Paperwork and hardware	50,000

Table 4: Budget of this study

CHAPTER 4
RESULT

4.1 Baseline characteristics

4.1.1 Baseline characteristics of non-time-series data

A total of 581 dialysis sessions from KCMH and 570 dialysis sessions from BKIH were included. Characteristics of dialysis sessions were categorized in two ways for baseline comparisons: by institute (KCMH and BKIH) and by model development process (training, validation, and test sets). Table 5 presents the relevant clinical data for subjects from KCMH and BKIH. While the two groups had comparable ages (64.9 vs. 67.7 years), patients from KCMH were more likely to be female (62.7% vs. 52.6%),

and had higher proportions of cerebrovascular disease, gout, and chronic lung disease. In contrast, BKIH had higher proportions of hypertension and diabetes (93.3% vs. 91.2% and 41.4% vs. 34.1%, respectively). Hemodiafiltration was the dominant mode of dialysis in KCMH, whereas hemodialysis was more prevalent in BKIH. Average DW from BCM was slightly lower in KCMH (54.9 kg vs. 5881 kg), while average of net ultrafiltration volume and average ultrafiltration rate was comparable in both groups (2.1 L vs 2.2 L) and (9.1 mL/kg/hr vs 9.7 L/kg/hr) respectively. Mean values of dialysate and laboratory data were comparable in both groups. Baseline characteristics of dialysis sessions by model development process was shown in table 6 and table 7. Figure 9 showed snippet of data in this domain.

	KCMH (n = 581 sessions)	BKIH (n = 570 sessions)
Number of patients	42	244
Age, mean (±SD) years	64.9 (9.6)	67.7 (12.3)
Female, number (%)	364 (62.7%)	300 (52.6%)
BCM-DW, mean (±SD) kg	54.9 (10.5)	58.8 (11.7)
Average Pre-dialysis BW, mean (±SD) kg	56.9 (10.6)	61.5 (11.7)
Average Post-dialysis BW, mean (±SD) kg	54.7 (10.3)	59.3 (11.6)
Average Height, mean (±SD) cm	159.0 (8.5)	160.8 (8.3)
Average BMI, mean (±SD) kg/m2	21.7 (3.8)	22.7 (4.0)
Net UF, mean (±SD) L	2.1 (0.9)	2.2 (0.8)
Average UFR, mean (±SD) mL/kg/hr	9.1 (4.4)	9.7 (4.0)
Comorbidities, number (%)		
Hypertension	530 (91.2%)	532 (93.3%)
Diabetes	198 (34.1%)	236 (41.4%)
Cerebrovascular disease	67 (11.5%)	29 (5.1%)
Gout	57 (9.8%)	47 (8.2%)
Chronic lung disease	28 (4.8%)	3 (0.5%)

Mode of dialysis, number (%)		
Hemodialysis	35 (6.0%)	570 (100%)
Hemodiafiltration	546 (94.0%)	0 (0%)
Dialysate profile data		
Dialysate Sodium, mean (±SD) mg/dL Median (±IQR)	137.4 (1.4) 138.0 (136.0 - 138.0)	137.7 (1.1) 138.0 (138.0 - 138.0)
Dialysate Bicarbonate, mean (±SD) mg/dL Median (±IQR)	32.0 (1.0) 32.0 (32.0 - 32.0)	32.1 (0.8) 32.0 (32.0 - 32.0)
Dialysate Potassium, mean (±SD) mg/dL Median (±IQR)	2.1 (0.3) 2.0 (2.0 - 2.0)	2.6 (0.5) 3.0 (2.0 - 3.0)
Dialysate Calcium, mean (±SD) mg/dL Median (±IQR)	3.0 (0.3) 3.0 (2.5 - 3.0)	2.8 (0.3) 3.0 (2.5 - 3.0)
Dialysate Temperature, mean (±SD) Celsius Median (±IQR)	36.8 (0.5) 37.0 (37.0 - 37.0)	36.5 (0.5) 36.5 (36.3 - 36.8)
Laboratory data		
Hemoglobin, mean (±SD) mg/dL	11.0 (1.4)	10.7 (1.5)
Calcium, mean (±SD) mg/dL	8.7 (0.7)	8.8 (0.8)
Sodium, mean (±SD) mg/dL	138.7 (2.7)	136.5 (4.2)
Albumin, mean (±SD) mg/dL	3.8 (0.3)	3.9 (0.4)

Table 5: Baseline characteristics of dialysis sessions from each institute.
Abbreviations: KCMH, King Chulalongkorn Memorial Hospital; BKIH, Bhumirajanagarindra Kidney Institute Hospital; BCM-DW, dry weight from BCM; BW, body weight; BMI, body mass index; SD, standard deviation; IQR, interquartile range

	KCMH (n = 581 sessions)		BKIH (n = 570 sessions)
	Training set	Validation set	Testing set

Number of sessions (%)	463 (79.7%)	118 (20.3%)	570
Number of patients	35	7	244
Age, mean (±SD) years	63.8 (9.9)	62.3 (6.5)	67.7 (12.3)
Female, number (%)	255 (55.1%)	109 (92.8%)	300 (52.6%)
BCM-DW, mean (±SD) kg	54.3 (10.9)	56.9 (10.2)	58.8 (11.7)
Average Pre-dialysis BW, mean (±SD) kg	56.4 (10.8)	58.9 (11.3)	61.5 (11.7)
Average Post-dialysis BW, mean (±SD) kg	54.1 (10.6)	56.8 (11.1)	59.3 (11.6)
Average Height, mean (±SD) cm	160.5 (8.3)	153.5 (7.1)	160.8 (8.3)
Average BMI, mean (±SD) kg/m2	21.0 (3.6)	24.1 (3.7)	22.7 (4.0)
Net UF, mean (±SD) L	2.1 (1.0)	2.0 (1.0)	2.2 (0.8)
Average UFR, mean (±SD) mL/kg/hr	9.0 (4.6)	9.3 (4.2)	9.7 (4.0)
Comorbidities, number (%)			
Hypertension	412 (89.0%)	118 (100%)	532 (93.3%)
Diabetes	115 (24.8%)	83 (70.3%)	236 (41.4%)
Cerebrovascular disease	16 (3.5%)	51 (43.2%)	29 (5.1%)
Gout	38 (8.2%)	19 (16.1%)	47 (8.2%)
Chronic lung disease	28 (6.0%)	0 (0%)	3 (0.5%)
Mode of dialysis, number (%)			
Hemodialysis	0 (0%)	35 (29.7%)	570 (100%)
Hemodiafiltration	463 (100%)	83 (70.3%)	0 (0%)
Dialysate profile data			
Dialysate Sodium, mean (±SD) mg/dL	137.2 (1.4)	138.3 (0.7)	137.7 (1.1)
Dialysate Bicarbonate, mean (±SD) mg/dL	31.9 (1.0)	32.2 (1.2)	32.1 (0.8)
Dialysate Potassium, mean (±SD) mg/dL	2.2 (0.4)	2.0 (0)	2.6 (0.5)
Dialysate Calcium, mean (±SD)	2.9 (0.3)	3.1 (0.2)	2.8 (0.3)

mg/dL			
Dialysate Temperature, mean (±SD) Celsius	36.8 (0.4)	36.6 (0.5)	36.5 (0.5)
Laboratory data			
Hemoglobin, mean (±SD) mg/dL	10.9 (1.4)	11.2 (1.2)	10.7 (1.5)
Calcium, mean (±SD) mg/dL	8.8 (0.7)	8.5 (0.7)	8.8 (0.8)
Sodium, mean (±SD) mg/dL	138.5 (2.7)	139.3 (2.2)	136.5 (4.2)
Albumin, mean (±SD) mg/dL	3.9 (0.3)	3.8 (0.3)	3.9 (0.4)

Table 6: Baseline characteristics of dialysis session by training and testing process (Configuration A).

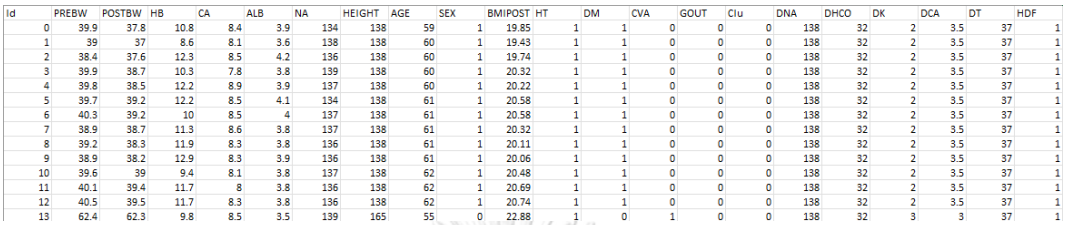
Abbreviations: KCMH, King Chulalongkorn Memorial Hospital; BKIH, Bhumirajanagarindra Kidney Institute Hospital; BCM-DW, dry weight from BCM; BW, body weight; BMI, body mass index; SD, standard deviation; IQR, interquartile range

	BKIH (n = 570 sessions)		KCMH (n = 581 sessions)
	Training set	Validation set	Testing set
Number of sessions (%)	455 (79.8%)	115 (20.1%)	581
Number of patients	200	44	42
Age, mean (±SD) years	63.8 (14.1)	70.1 (10.2)	64.9 (9.60)
Female, number (%)	221 (48.6%)	79 (68.7%)	364 (62.7%)
BCM-DW, mean (±SD) kg	61.3 (11.1)	59.1 (12.9)	54.9 (10.5)
Average Pre-dialysis BW, mean (±SD) kg	64.5 (10.9)	61.8 (12.9)	56.9 (10.6)
Average Post-dialysis BW, mean (±SD) kg	62.0 (10.9)	59.6 (12.7)	54.7 (10.3)
Average Height, mean (±SD) cm	165.6 (7.8)	159.7 (7.6)	159.0 (8.5)
Average BMI, mean (±SD) kg/m2	22.3 (3.6)	23.1 (4.5)	21.7 (3.8)
Net UF, mean (±SD) L	2.2 (0.9)	2.1 (0.9)	2.1 (0.9)

Average UFR, mean (±SD) mL/kg/hr	9.8 (4.1)	9.6 (4.1)	9.1 (4.4)
Comorbidities, number (%)			
Hypertension	420 (92.3%)	112 (97.4%)	530 (91.2%)
Diabetes	177 (38.9%)	59 (51.3%)	198 (34.1%)
Cerebrovascular disease	22 (4.8%)	7 (6.1%)	67 (11.5%)
Gout	31 (6.8%)	16 (13.9%)	57 (9.8%)
Chronic lung disease	1 (0.2%)	2 (1.7%)	28 (4.8%)
Mode of dialysis, number (%)			
Hemodialysis	455 (100%)	115 (100%)	35 (6.0%)
Hemodiafiltration	0 (0%)	0 (0%)	546 (94.0%)
Dialysate profile data			
Dialysate Sodium, mean (±SD) mg/dL	137.8 (1.0)	137.4 (1.6)	137.4 (1.4)
Dialysate Bicarbonate, mean (±SD) mg/dL	32.1 (1.1)	32.2 (1.4)	32.0 (1.0)
Dialysate Potassium, mean (±SD) mg/dL	2.9 (0.4)	2.2 (0.5)	2.1 (0.3)
Dialysate Calcium, mean (±SD) mg/dL	2.9 (0.2)	2.7 (0.4)	3.0 (0.3)
Dialysate Temperature, mean (±SD) Celsius	36.5 (0.4)	36.3 (0.6)	36.8 (0.5)
Laboratory data			
Hemoglobin, mean (±SD) mg/dL	10.5 (1.3)	10.8 (1.4)	11.0 (1.4)
Calcium, mean (±SD) mg/dL	8.9 (0.7)	8.8 (0.9)	8.7 (0.7)
Sodium, mean (±SD) mg/dL	135.7 (4.4)	137.9 (3.1)	138.7 (2.7)
Albumin, mean (±SD) mg/dL	3.9 (0.5)	3.9 (0.3)	3.8 (0.3)

Table 7: Baseline characteristics of dialysis session by training and testing process (Configuration B)

Abbreviations: KCMH, King Chulalongkorn Memorial Hospital; BKIH, Bhumirajanagarindra Kidney Institute Hospital; BCM-DW, dry weight from BCM; BW, body weight; BMI, body mass index; SD, standard deviation; IQR, interquartile range



Id	PREBW	POSTBW	HB	CA	ALB	NA	HEIGHT	AGE	SEX	BMIPOST	HT	DM	CVA	GOUT	CIu	DNA	DHCO	DK	DCA	DT	HDF
0	39.9	37.8	10.8	8.4	3.9	134	138	59	1	19.85	1	1	0	0	0	138	32	2	3.5	37	1
1	39	37	8.6	8.1	3.6	138	138	60	1	19.43	1	1	0	0	0	138	32	2	3.5	37	1
2	38.4	37.6	12.3	8.5	4.2	136	138	60	1	19.74	1	1	0	0	0	138	32	2	3.5	37	1
3	39.9	38.7	10.3	7.8	3.8	139	138	60	1	20.32	1	1	0	0	0	138	32	2	3.5	37	1
4	39.8	38.5	12.2	8.9	3.9	137	138	60	1	20.22	1	1	0	0	0	138	32	2	3.5	37	1
5	39.7	39.2	12.2	8.5	4.1	134	138	61	1	20.58	1	1	0	0	0	138	32	2	3.5	37	1
6	40.3	39.2	10	8.5	4	137	138	61	1	20.58	1	1	0	0	0	138	32	2	3.5	37	1
7	38.9	38.7	11.3	8.6	3.8	137	138	61	1	20.32	1	1	0	0	0	138	32	2	3.5	37	1
8	39.2	38.3	11.9	8.3	3.8	136	138	61	1	20.11	1	1	0	0	0	138	32	2	3.5	37	1
9	38.9	38.2	12.9	8.3	3.9	136	138	61	1	20.06	1	1	0	0	0	138	32	2	3.5	37	1
10	39.6	39	9.4	8.1	3.8	137	138	62	1	20.48	1	1	0	0	0	138	32	2	3.5	37	1
11	40.1	39.4	11.7	8	3.8	136	138	62	1	20.69	1	1	0	0	0	138	32	2	3.5	37	1
12	40.5	39.5	11.7	8.3	3.8	136	138	62	1	20.74	1	1	0	0	0	138	32	2	3.5	37	1
13	62.4	62.3	9.8	8.5	3.5	139	165	55	0	22.88	1	0	1	0	0	138	32	3	3	37	1

Figure 9: A snippet of the laboratory, demographic, and dialysis data.
Abbreviations: Id, identification number of patient; PREBW, pre, systolic blood pressure; DBP, diastolic blood pressure; MAP, mean arterial pressure; HR, heart rate; UFR, ultrafiltration rate

4.1.2 Baseline characteristics of time-series data

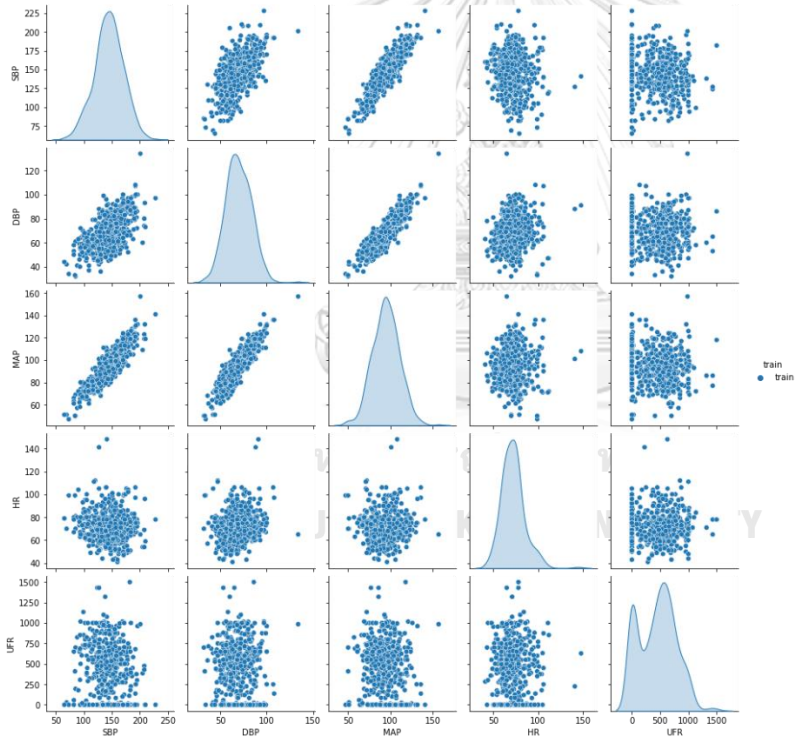
Descriptive statistics of time-series data from overall dialysis session in each institute was shown in table 6. The values of each parameter were quite similar, although BKIH group had a little higher mean of systolic blood pressure and ultrafiltration rate. For easier visualization, the relationship between each feature of time-series data was shown in Figure 10. Specifically, change of heart rate during dialysis session was shown in the spaghetti plot of Figure 11. A part of data was shown in snippet in Figure 12.

	KCMH (n = 581 sessions)			BKIH (n = 570 sessions)		
	Max	Min	Mean (SD)	Max	Min	Mean (SD)
SBP (mmHg)	238	52	145.00 (25.85)	240	61	149.75 (23.04)

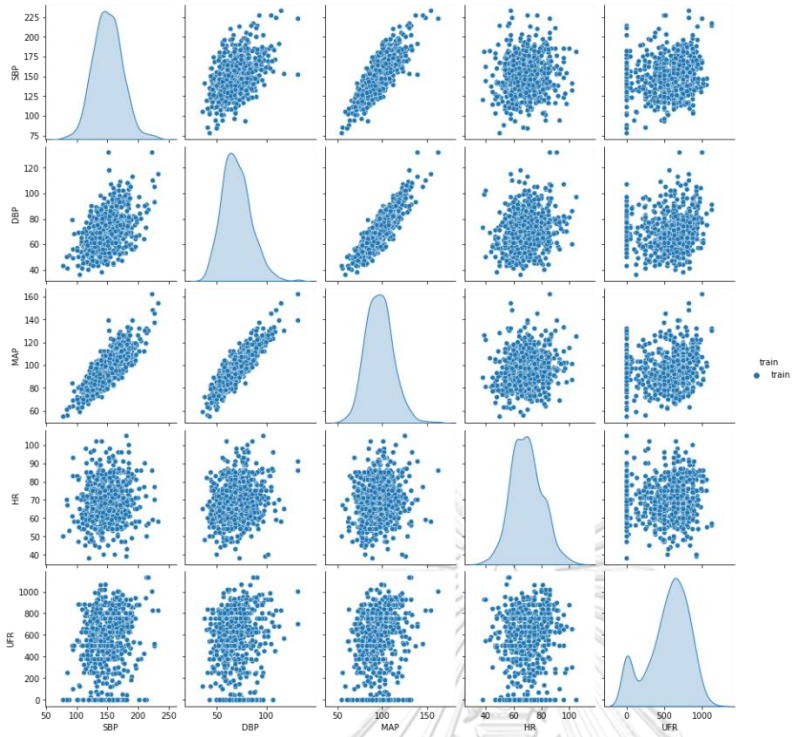
DBP (mmHg)	154	22	68.65 (13.55)	162	33	69.75 (14.42)
MAP (mmHg)	159	44	94.43 (15.42)	169	49	96.11 (14.88)
HR (bpm)	180	30	71.24 (13.17)	145	37	69.50 (11.89)
UFR (ml/min)	1650	0	450.71 (313.07)	3000	0	541.90 (270.18)

Table 8: Descriptive statistics of time-series data.

Abbreviations: KCMH, King Chulalongkorn Memorial Hospital; BKIH, Bhumirajanagarindra Kidney Institute Hospital; maximum, Max; minimum, Min; standard deviation, SD; SBP, systolic blood pressure; DBP, diastolic blood pressure; MAP, mean arterial pressure; HR, heart rate; UFR, ultrafiltration rate



KCMH



BKIH

Figure 10: A scatter plot showed the relationship between each feature of the time-series data was displayed. KCMH (upper) and BKIH (lower)

Abbreviations: SBP, systolic blood pressure; DBP, diastolic blood pressure; MAP, mean arterial pressure; HR, heart rate; UFR, ultrafiltration rate

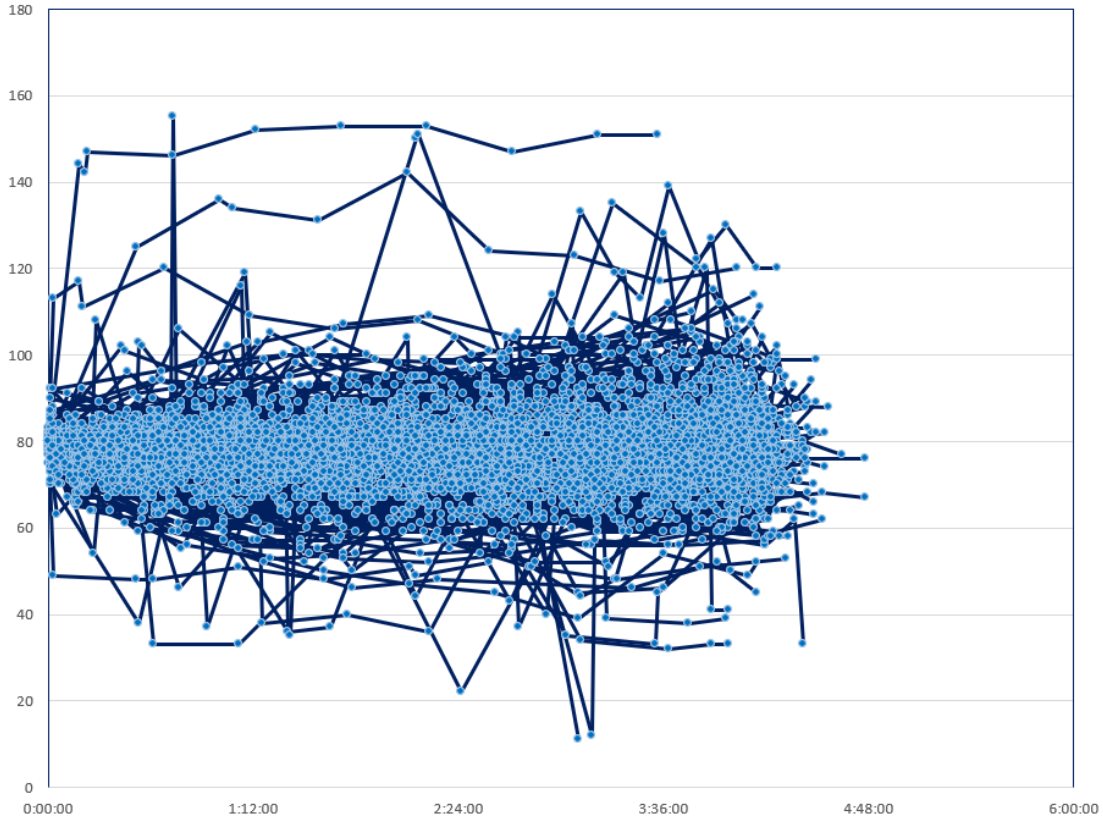


Figure 11: Spaghetti plot showed HR variation over time in each dialysis session.

patient code	Dialysis session	Time	Number of timestamp	SBP (mmHg)	DBP (mmHg)	MAP (mmHg)	HR (bpm)	UFR (ml/min)
1	1	05.02	0	162	57	92	75	725
1	1	05.31	1	150	58	89	73	797
1	1	06.03	2	134	52	80	77	725
1	1	07.02	3	135	61	86	76	725
1	1	08.02	4	138	57	84	71	532
1	1	09.10	5	99	82	88	65	5
1	1	09.32	6	113	46	69	65	5

Figure 12: A snippet of the time-series data in one session was shown.

Abbreviations: SBP, systolic blood pressure; DBP, diastolic blood pressure; MAP, mean arterial pressure; HR, heart rate; UFR, ultrafiltration rate

4.1.3 Distribution of output (BCM-DW)

The target for model's learning in this project is BCM-DW. The histogram of BKIH showed normal distribution, while KCMH database showed a minute right-skewed distribution.

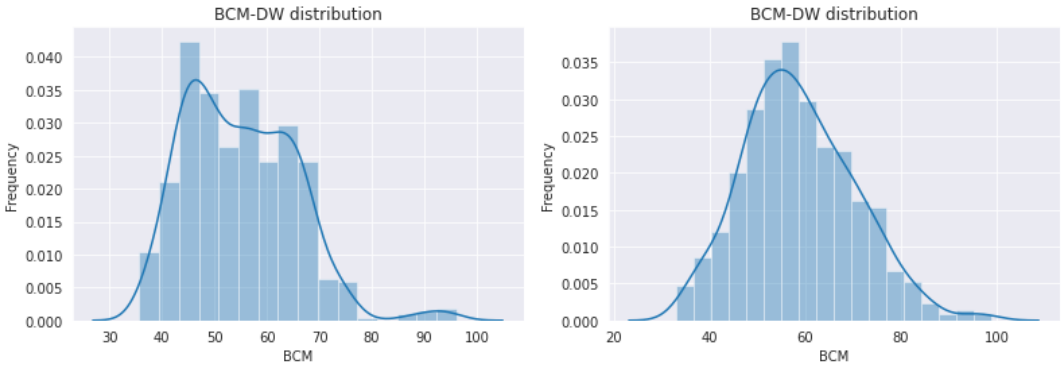


Figure 13: Histogram of the BCM-DW (kg) of each dialysis session. KCMH (left) and BKIH (right).

Abbreviations: KCMH, King Chulalongkorn Memorial Hospital; BKIH, Bhumirajanagarindra Kidney Institute Hospital; BCM, Body composition monitor; BCM-DW, Dry weight from Body composition monitor

4.2 Result from model development phase

4.2.1 Processing of time-series data

4.2.1.1 LSTM as regression model for DW prediction (LSTMreg)

In this approach, the LSTMreg was not able to effectively predict DW based on the time-series data, as evidenced by the high MSE values of 151.3991 (SD 4.3234) and 125.8074 (SD 2.3913) for configurations A and B, respectively. Additionally, poor agreement was observed, as shown in figure 14, indicating that the model did not accurately capture the patterns and trends in the data. Therefore, the LSTM model is not a suitable regression model for DW prediction in this case.

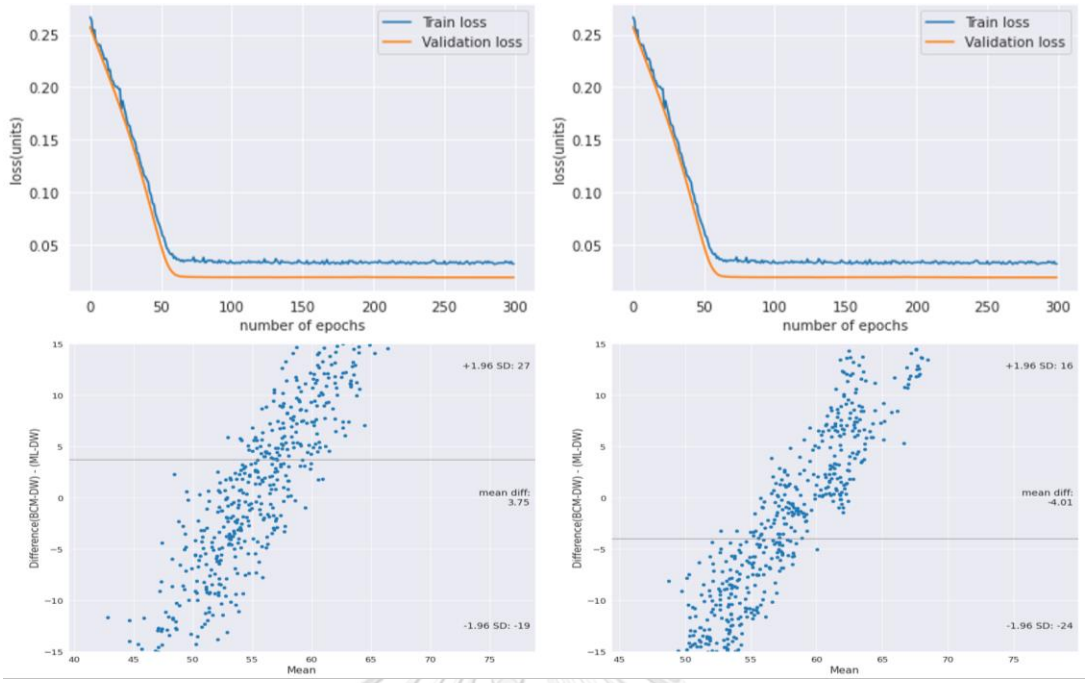


Figure 14: Training and validation graph loss (upper), Bland-Altman plot of prediction from LSTM and BCM-DW (lower). Image showed result of LSTM regression on both configuration A (left) and configuration B (right).

4.2.1.2 LSTM as classification model for grouping dialysis session (LSTMclass)

In the classification task, the LSTMclass was able to categorize dialysis sessions into groups of 0, 1, and 2 based on different net UF values from small to large, respectively. However, the accuracy of the classification was poor, with only 48.07% in configuration A and 61.45% in configuration B. The precision was 59.79% and 46.25%, respectively, and the F1 score showed 46.96% and 52.75%, respectively. Although the confusion matrix had a diagonal shape, it was not perfectly aligned. The overall accuracy rate of 50% was also not sufficient to use this as an initial model, even though a few subgroups were correctly classified. To determine the potential of the model in separating these groups, unsupervised learning of the last embedding layers was conducted. The Haversine plot in Figure 16 demonstrated that the model had the capability to separate samples of each subgroup into distinct zones, even without explicit supervision.

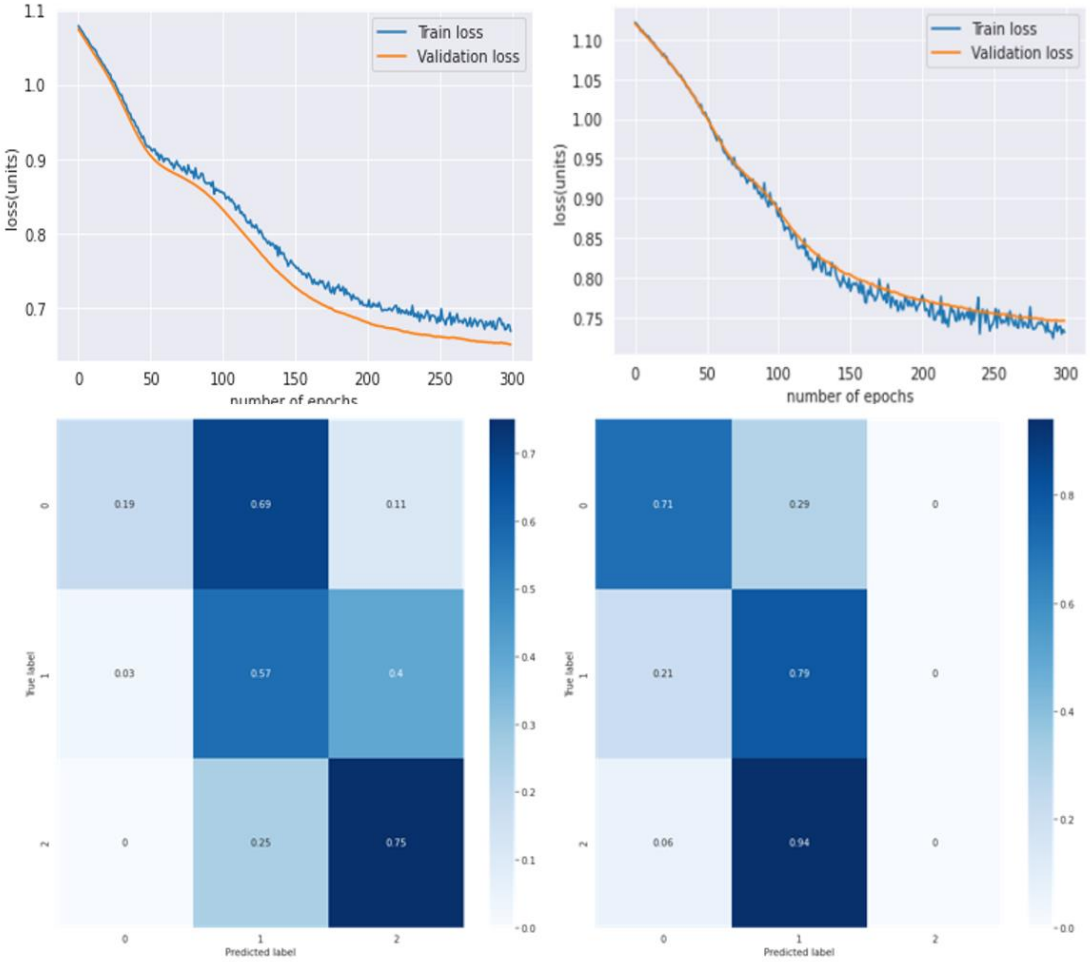


Figure 15: Training and validation graph loss (upper). Confusion matrix of grouping of dialysis session categorized by net UF. Image showed result of LSTM on both configuration A (left) and configuration B (right)

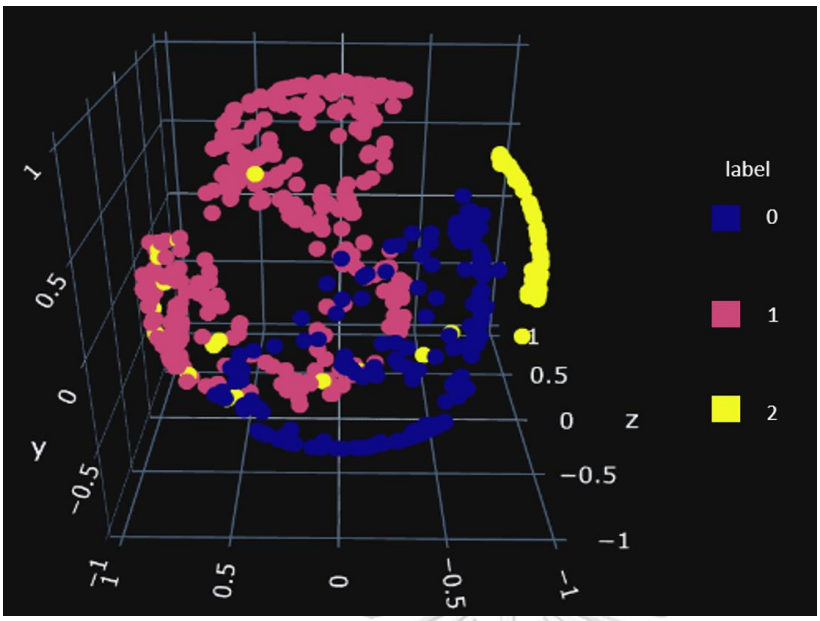


Figure 16: UMAP with Haversine plot showed unsupervised dimension reduction, with the input data being the embedding layers from LSTM classification.

4.2.2 Processing of non-time-series data

4.2.2.1 multiple linear regression (ML)

In this approach, MSE in configuration A was 5.7497, meanwhile configuration B showed 3.1051. In Bland-Altman plot between this model and BCM-DW showed mean of different at -1.47 kg in configuration A and +0.27 kg in configuration B. While limit of agreement was up to 5.2 kg.

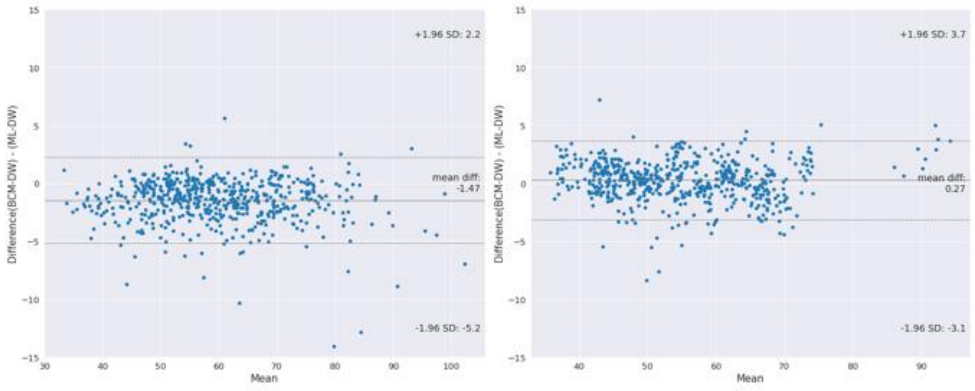


Figure 17: Bland-Altman plot of multiple linear regression model and BCM-DW. configuration A (left) and configuration B (right).

4.2.2.2 Machine learning models (STACK)

The machine learning approach showed lower MSE in every base model, as shown in Table 7. The Bland-Altman plot between this method and BCM-DW (figure 17) showed that the mean difference was -0.78 kg in configuration A and +0.02 kg in configuration B. The stacked model was comprised of Lasso, ENet, KRR and LightGBM models.

	MSE	
	Configuration A	Configuration B
Lasso model	1.6187	1.7609
ENet model	1.6187	1.7609
KRR model	2.3424	1.5338
XGBoost model	4.8185	3.0601
LightGBM model	2.8381	1.8436
Stacked model	1.6898	1.7867

Table 9: MSE of each ML model in both configuration

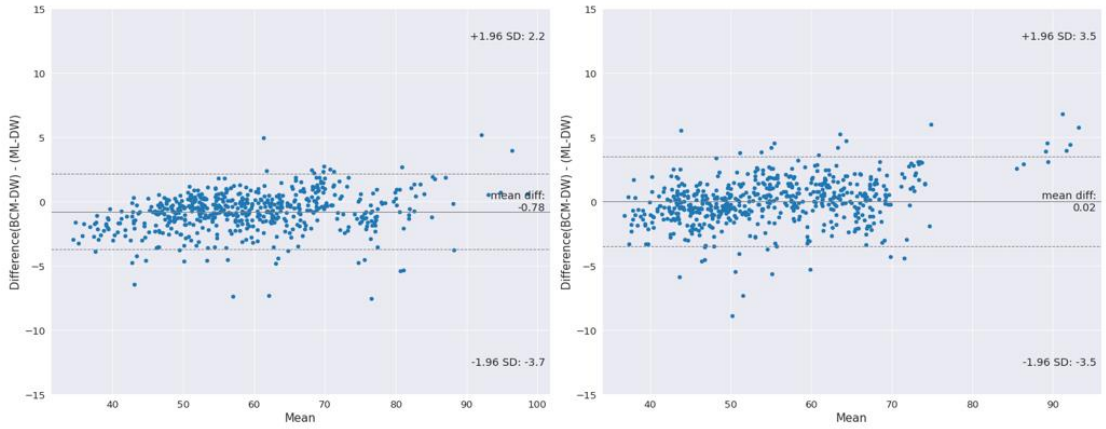


Figure 18: Bland-Altman plot between STACK ML model and BCM-DW. Configuration A (left) and configuration B (right).

4.2.2.3 Neural network model (gNN)

Approach by neural network for general data, the model showed an MSE of 5.7527 (SD 2.0695) in configuration A and 5.1071 (SD 4.0361) in configuration B. Bland-Altman plots between this method and BCM-DW in both configurations are shown in Figure 18.

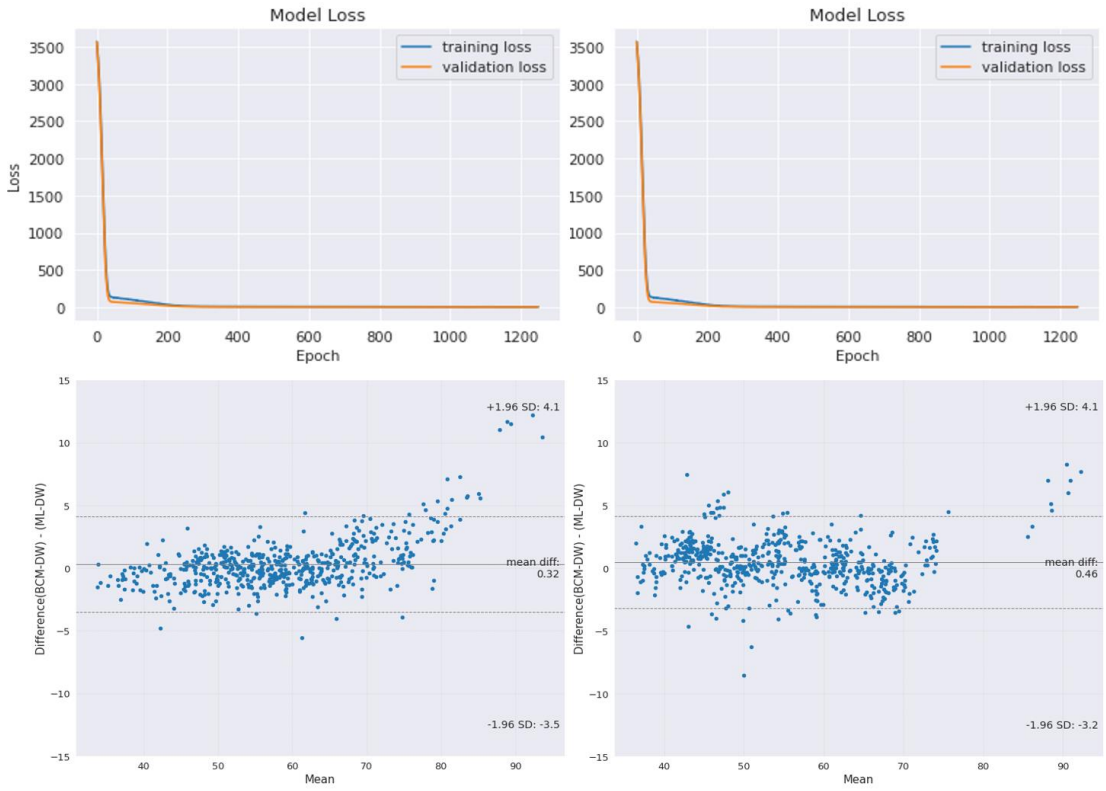


Figure 19: Bland-Altman plot between non-time-series NN model and BCM-DW. Configuration A (left) and configuration B (right).

4.2.3 Combined result of time-series data and non-time-series data (gtNN)

In this approach, LSTM was used as a base model, and non-time-series data embedding was combined before sending data through the last dense layer. The

optimization was done for both types of data simultaneously. However, the model showed a high MSE of 15.3955 (SD 11.3553) in configuration A and 7.8782 (SD 3.9838) in configuration B. The results of the Bland-Altman plot in Figure 19 showed that the mean difference was -1.24 kg and +0.49 kg in configuration A and configuration B, respectively. The limit of agreement was up to 7.7 kg in configuration A and 5.6 kg in configuration B, indicating a wide range of error.

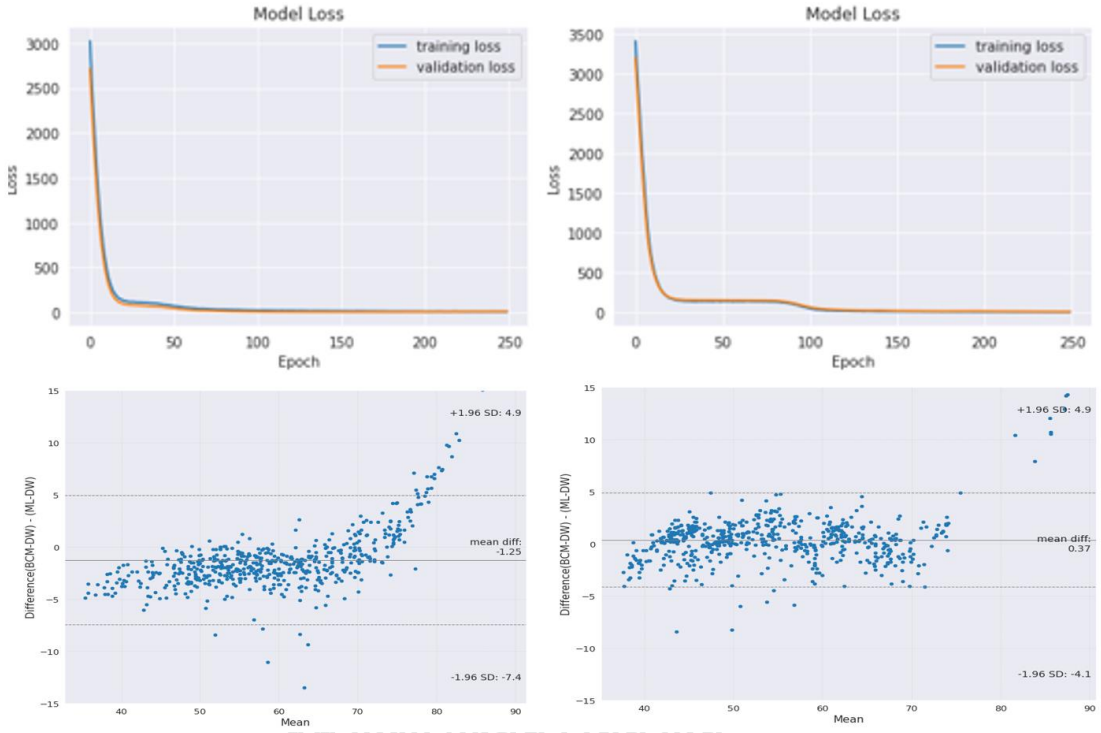
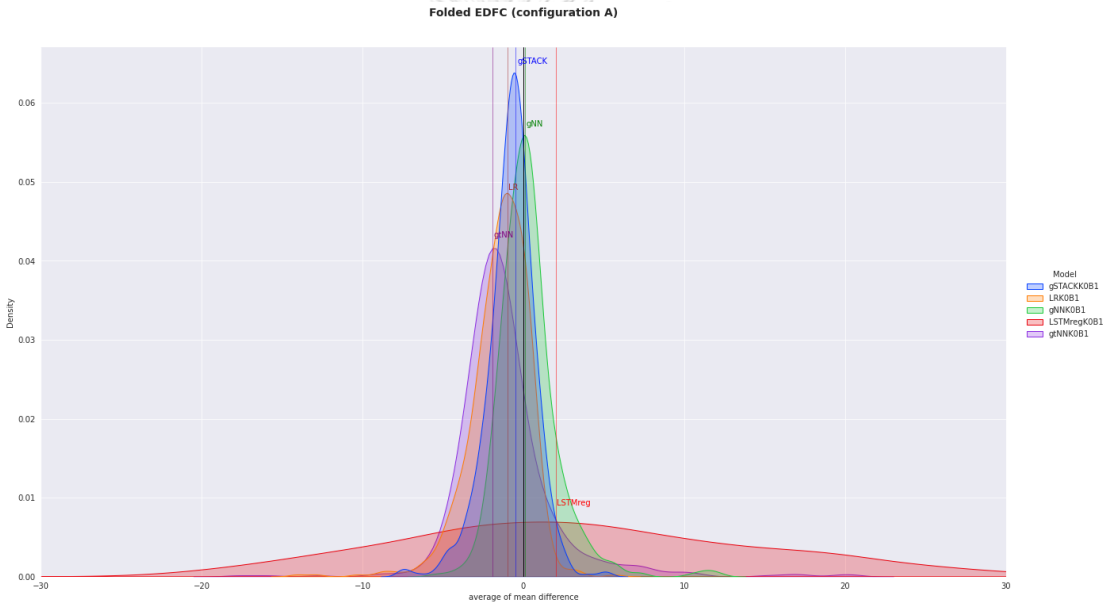


Figure 20: Bland-Altman plot between NN model of combined embedding layer from both time-series data and non-time-series data and BCM-DW. Configuration A (left) and configuration B (right).

Based on the results of all approaches, it appears that using time-series data alone was not a good input for predicting BCM-DW. On the other hand, the model for non-time-series data was able to predict BCM-DW with better accuracy. However, when time-series data was combined with it, the performance decreased. The comparison of the Bland-Altman plot between the different approaches showed

slight differences, with the limit of agreement being wider for the time-series data models.

The folded EDFC was also plotted to evaluate the results from the prediction of each model (figure 20). In both configurations, the graph from the LSTMreg model showed the lowest peak and the widest range of sway from the actual value. On the other hand, the graph from the STACK model gave the nearest peak to the zero line and was the steepest among the other graphs. Therefore, the STACK model was chosen as the final model for this study.



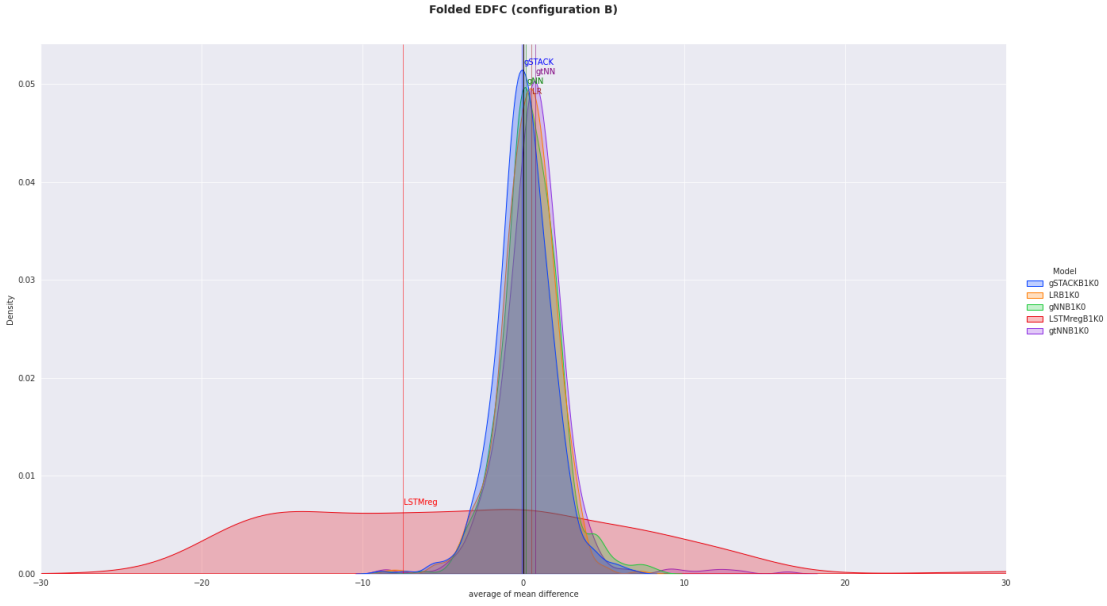


Figure 21: Folded Empirical Distribution Function Curves of each model. Configuration A (upper) and Configuration B (lower).

4.3 Result from further analysis of final model

Based on the presented results, the final model for predicting dry weight (DW) showed a mean difference of -0.78 kg and +0.02 kg when tested in configuration A and configuration B, respectively. The absolute mean difference was 1.25 kg and 1.32 kg for configuration A and configuration B, respectively. The limits of agreement were wider for configuration A (-3.7 to +2.2 kg) compared to configuration B (-3.5 kg to +3.5 kg). The subgroup above the limit contained 11 samples in configuration A and 16 samples in configuration B, while the subgroup below the limit contained 20 samples in configuration A and 13 samples in configuration B. In summary, the number of samples out of agreement was 31 out of 570 samples (5.44%) and 29 out of 581 samples (4.99%) for configurations A and B, respectively. It is worth noting that these sessions that caused an outlier in prediction did not come from a single specific patient but were spread across 15 patients in the KCMH database and 25 patients in the BKIH database. These results suggest that while the final model may have some degree of predictive accuracy, there is still room for improvement in terms of reducing variability and increasing agreement between predicted and actual values.

This was done by dividing the samples into three subgroups based on the range of agreement, as shown in Figure 21. Two sub-analyses were then performed to determine the cause of poor performance in some samples.

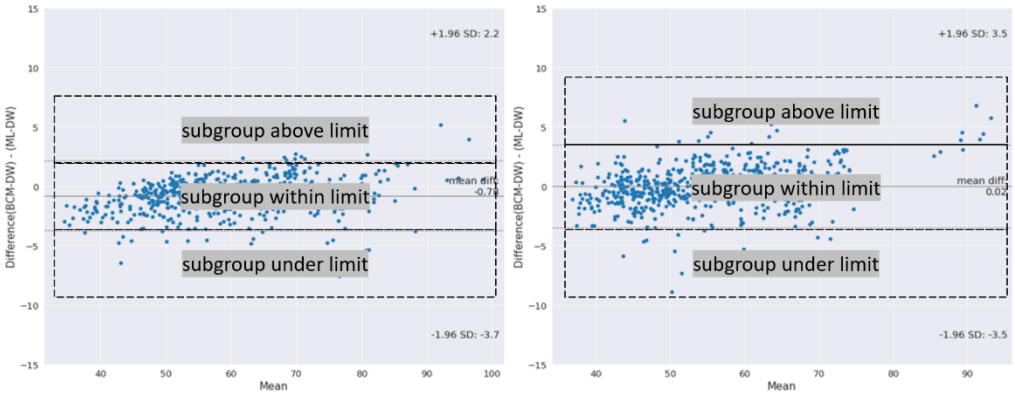


Figure 22: Divided dialysis group of final model based on limit of agreement. Configuration A (left) and Configuration B (right).

The sub-analysis using SHAP values to determine the effect of each parameter on the model's prediction was shown in Figure 22. This can be particularly useful in cases where the model has a large number of features, as it can help to reduce the computational resources required for training and inference by eliminating unnecessary features. However, in this study, we aimed whether certain features were driving the results into different subgroups.

The second thing that assist this sub-analysis was reviewing the descriptive data of outliers to see whether interested feature with abnormal SHAP value did have an abnormal distribution between the subgroups (Table 10 and Table 11). This, combined with the study of SHAP values, can help identify at risk features that may be contributing to the poor performance of the model. If such features were identified, we would evaluate whether removing them could lead to fewer instances of disagreement.

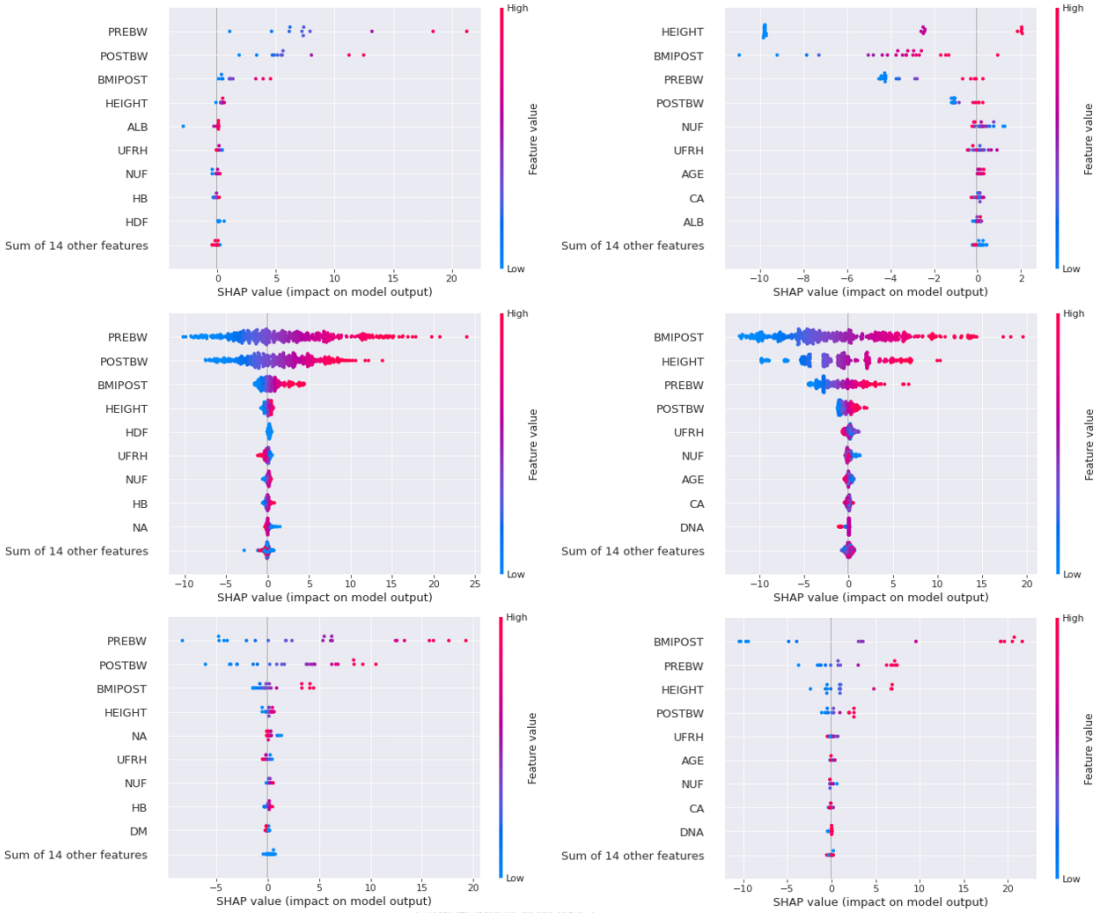


Figure 23: SHAP value of each subgroup in each configuration. Subgroup above limit (upper), Subgroup within limit (middle), Subgroup below limit (lower). Configuration A (left) and Configuration B (right)

	above LOA			Within LOA			Below LOA		
	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
PREBW	75.3	100	59.5	61	106.3	35.7	65.4	93.5	40.8
POSTBW	73.5	97.1	59.3	58.8	102.7	33.1	62.9	91.9	38.3
NUF	1.7	2.9	0.2	2.2	4.2	0	2.5	3.9	1.4
UFRH	5.9	11.3	0.8	9.7	24.9	0	10.7	16.7	4.1
HB	9.9	12.3	8.4	10.7	16.7	6.5	10.8	13.3	8.9
CA	8.8	9.5	7.7	8.8	14.2	5.4	9.1	10.9	7.8
ALB	3.6	4.2	1.5	3.9	4.9	1.5	3.9	4.7	3.2
NA	136.7	141	129	136.6	146	114	135	143	120
HEIGHT	168.4	176	156	160.6	183	141	162	176	148
AGE	67.2	86	37	67.8	89	31	62.7	77	46

SEX	0.1	1	0	0.5	1	0	0.5	1	0
BMIPOST	26.4	34	22.3	22.6	34	15	21.9	32.6	15.7
HT	1	1	1	0.9	1	0	0.7	1	0
DM	0.3	1	0	0.4	1	0	0.5	1	0
CVA	0	0	0	0	1	0	0	0	0
GOUT	0	1	0	0	1	0	0.1	1	0
Clu	0	0	0	0	1	0	0	0	0
DNA	138	138	138	137.6	140	130	138.1	140	138
DHCO	32	32	32	32	38	26	32	32	32
DK	2.8	3	2	2.5	4	2	2.7	3	2
DCA	2.8	3.5	2	2.7	3.5	2	2.8	3	2.5
DT	36.3	37.2	35.9	36.4	37.8	34.3	36.5	37.3	35
HDF	0	0	0	0	0	0	0	0	0

Table 10: Review of descriptive data between subgroups (configuration A)
Abbreviations: LOA, limit of agreement; PREBW, pre-dialytic body weight; POSTBW, post-dialytic body weight; NUF, net ultrafiltration volume; UFRH, ultrafiltration rate per hour; HB, hemoglobin; CA, calcium; ALB, albumin; NA, sodium; HT, hypertension; DM, diabetes; CVA, cerebrovascular accident; Clu, chronic lung disease; DNA, dialysate sodium; DHCO, dialysate bicarbonate; DK, dialysate potassium; DCA, dialysate calcium; DT, dialysate temperature; HDF, hemodiafiltration

	Above LOA			Within LOA			Below LOA		
	mean	max	min	mean	max	Min	Mean	Max	min
PREBW	72	96.7	42.4	56.4	91.5	38.4	57.1	75.5	46.9
POSTBW	69.4	93.3	40.4	54.3	89.7	36.7	55.3	73.8	45.7
NUF	2.5	3.6	0.6	2	4.5	0	1.8	3.9	-0.6
UFRH	9.7	13.6	1.6	9.7	24.4	0	8.3	21.3	-3.1
HB	10.9	13.9	7.8	10.9	15.6	6.2	10.7	12.7	8.7
CA	8.7	9.4	7.8	8.7	11.8	7	8.7	10.1	7
ALB	3.7	4.2	2.9	3.8	4.8	2.5	3.5	4.5	2.3
NA	138.4	144	132	138.7	145	129	136	140	133
HEIGHT	164.5	176	155	158.6	183	138	166.7	176	141
AGE	60.3	67	50	65.1	83	39	58	76	52
SEX	0.5	1	0	0.6	1	0	0.1	1	0

BMIPOST	25.7	35.1	16.5	21.5	33.7	15.2	19.9	24.3	16.7
HT	0.8	1	0	0.9	1	0	0.9	1	0
DM	0	1	0	0.3	1	0	0.3	1	0
CVA	0	1	0	0.1	1	0	0	0	0
GOUT	0	1	0	0	1	0	0	1	0
Clu	0	1	0	0	1	0	0	0	0
DNA	137.6	138	136	137.3	140	134	136.6	138	134
DHCO	32	32	32	31.9	34	28	32.3	34	32
DK	2	2	2	2.1	3	2	2.2	3	2
DCA	3.1	3.5	3	2.9	3.5	2.5	3	3.5	2.5
DT	36.9	37	36	36.7	37	35	36.9	37	36.5
HDF	1	1	1	0.9	1	0	1	1	1

Table 11: Review of descriptive data between subgroups (configuration B)
Abbreviations: LOA, limit of agreement; PREBW, pre-dialytic body weight; POSTBW, post-dialytic body weight; NUF, net ultrafiltration volume; UFRH, ultrafiltration rate per hour; HB, hemoglobin; CA, calcium; ALB, albumin; NA, sodium; HT, hypertension; DM, diabetes; CVA, cerebrovascular accident; Clu, chronic lung disease; DNA, dialysate sodium; DHCO, dialysate bicarbonate; DK, dialysate potassium; DCA, dialysate calcium; DT, dialysate temperature; HDF, hemodiafiltration

In configuration A, SHAP values were calculated for each subgroup. However, the top four SHAP values were found to be the same across all subgroups, namely "Pre-dialytic BW," "Post-dialytic BW," "BMI," and "Height," respectively. This suggests that searching for a culprit feature that drives the results to be outliers may not be successful since there was no striking feature between the subgroups. Although there was a difference in the 5th rank of SHAP value between the subgroups, the effect of this rank was quite small. The subgroup within the limit had "HDF status" as their 5th rank, while the subgroup above the limit had "serum sodium," and the subgroup below the limit had "serum albumin," respectively. A review of the demographic data showed that these outliers did not come from the same patient. Further description in Table 8 showed that the average values of “serum sodium” between each

subgroup were not much different, while “serum albumin” varied at a greater amount. Therefore, the removal of this feature was expected to show more effect. Unfortunately, the result of the sensitivity analysis (figure 23) did not show much difference in the number of outliers (27 out of 570 samples, number out of agreement 4.74%). In conclusion, the results that cause outliers in configuration A may not be due to any single parameter, but rather due to the variation of raw data itself.

In configuration B, the top 4 SHAP values for each subgroup were the same, which were "Height," "BMI," "Pre-dialytic BW," and "Post-dialytic BW," but the order was different for each subgroup. The subgroup above the limit had "Height" as the first rank of SHAP value, while it was the second and third rank in the subgroup within the limit and subgroup below the limit, respectively. Removing this parameter was expected to decrease the number of outliers that come from the subgroup above the limit. However, the sensitivity analysis (Figure 24) showed that while the subgroup above the limit contained fewer outliers, decreasing from 16 to 12 samples, the subgroup below the limit contained more outliers, increasing from 13 to 18 samples. In summary, removing a single parameter could not simply decrease an outlier as it also affects other subgroups.

We were also concerned about the parameter "Post-dialytic BW," as it may be dependent on the output value, so we conducted a sub-analysis to see if the model relied solely on this parameter or not. The sub-analysis (Figure 25) showed that the mean difference changed slightly from +0.02 kg to -0.03 kg. The limit of agreement also increased slightly from a lower limit of -3.5 kg to -3.7 kg and a lower limit of +3.5 kg to +3.7 kg.

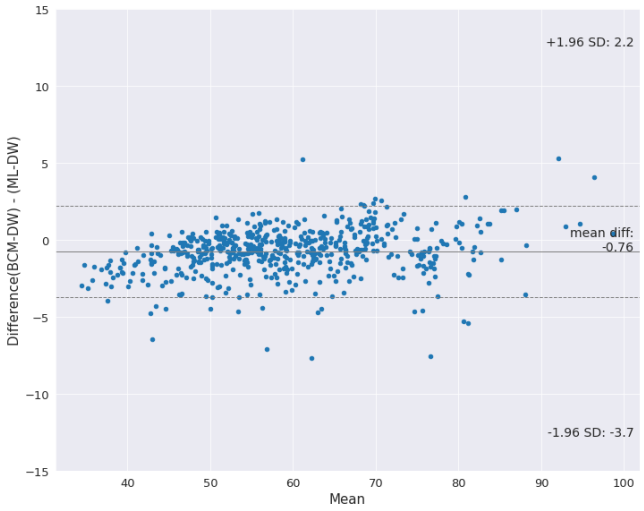


Figure 24: Bland-Altman plot of sub-analysis (exclude “serum albumin”) between ML-DW from STACK model and BCM-DW in Configuration A

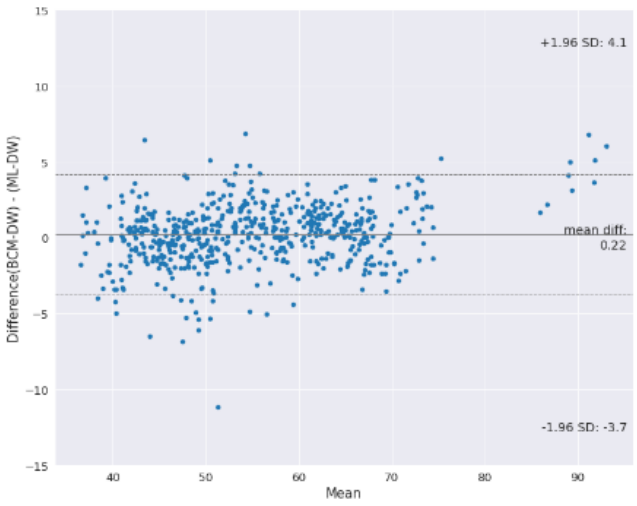


Figure 25: Bland-Altman plot of sub-analysis (exclude “Height”) between ML-DW from STACK model and BCM-DW in Configuration B

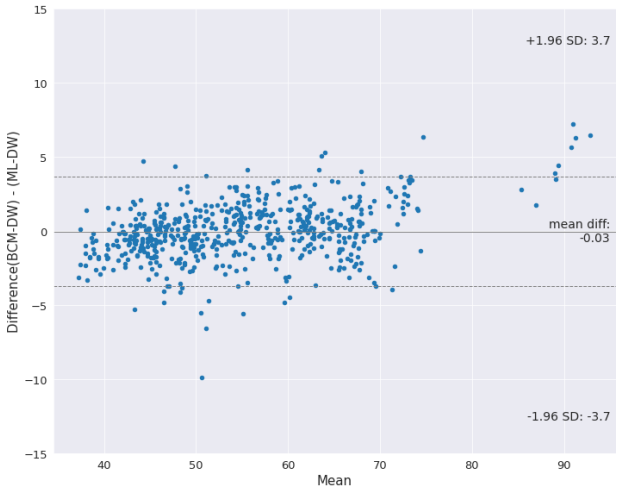


Figure 26: Bland-Altman plot of sub-analysis (exclude “Post-dialytic BW”) between ML-DW from STACK model and BCM-DW in Configuration B

Time latency for running this final model was 160 milliseconds. Size of the code was 420 KB.

CHATER 5
DISCUSSION

The present study aimed to predict BCM-DW using machine learning. The final model incorporated input variables mainly from demographic data, laboratory data, and dialysis data. The model of interest was a stacking of machine learning base models. The resulting model, ML-DW, was compared to BCM-DW using a test of agreement. Further sub-analysis was conducted to mitigate the effect of outliers. The robustness of the model was confirmed by testing it in a different database from another institute.

5.1 Descriptive analysis

The baseline characteristics of the dialysis sessions between the institutes were shown as described. Although this study focused on each individual dialysis session and the total number of dialysis sessions between the institutes was similar, the number of patients in the KCMH database was quite different from that in the BKIH database. This difference in number may imply that there were many dialysis sessions that contained demographic data from the same patients. This issue could have affected the model's development process. Therefore, we used the group shuffle split function to ensure that the sampling from the training and validation processes would not be contaminated by each other.

A concern about post-dialytic body weight in the KCMH database was also noted, as it was found to be quite close to BCM-DW. Therefore, any model that weighted the post-dialytic body weight parameter too heavily would make it easier for the model to roughly predict BCM-DW on KCMH data. However, this issue could be addressed by testing a model in both configuration A and B. Furthermore we aim for conducting a sub-analysis with the exclusion of the post-dialytic body weight parameter. Furthermore, this study aimed to test agreement, meaning that any small

variation between post-dialytic body weight and BCM-DW in each session would be taken into account during the assessment. Since post-dialytic body weight and BCM-DW have different standard deviations, a model that simply rounds the post-dialysis body weight to BCM-DW would not achieve good agreement.

Another important consideration in this study is the variation in time-series data. As shown in the scatter plot, there was no relationship between blood pressure and heart rate at all. Similarly, the trend of change in heart rate upon change in spaghetti plot was not observed, as shown in Figure 11. This is a concerning sign that these parameters may not provide much insight into the problem at hand, as there could be various factors that explain the lack of relationship.

Firstly, it should be noted that the vital sign and ultrafiltration rate trends in this study were adjusted preemptively. One of the institutes was a university hospital and the other a private hospital, and both had physicians rounding on patients in every dialysis session, even on holidays. They tended to make preemptive adjustments to prescriptions during dialysis in response to changes in intradialytic parameters aiming for minimize the adverse events, such as lowering the ultrafiltration rate when there was an alarm of blood volume monitoring to avoid consequences of intradialytic hypotension. Preemptive adjustments to prescriptions during dialysis likely mitigated changes in vital signs and reduced the influence of positive signals from vital signs and ultrafiltration rate on the determination of proper weight.

Secondly, despite preemptive efforts to minimize adverse events, they still occurred during dialysis sessions, such as hoarseness, cramping, and lightheadedness, which led to a reduction in the ultrafiltration rate. Thus, interruptions or changes in the ultrafiltration rate during dialysis could be a clue toward under dry weight. It is important to note that changes in ultrafiltration rate can also occur due to reasons other than adverse events. So, it needs to be clarify whether change in ultrafiltration were planned or unplanned and related to fluid status or not. However, records of

these events were not available and were not included in this study, thus limiting the ability to fully understand the impact of adverse events and ultrafiltration rate towards proper weight.

Thirdly, it should be noted that heart rate variability in this study might be compromised by beta blockers or sinus node dysfunction, which commonly found in senile adults. This factor would limit the potential usefulness of the designed system as it is based on tracking the HR change.

Finally, it should be mentioned that most patients in the university hospital used the hemodiafiltration technique, which is known to reduce the incidence of intradialytic hypotension. This could have weakened the effect of vital sign changes and reduced the accuracy of classification, even in patients who had high net UF. Although many approaches and models were tried in this study to figure out the usefulness of this time-series data, these factors should be taken into consideration when interpreting the results.

5.2 Model's development phase

Despite various attempts using different approaches, such as LSTMreg and LSTMclass, the LSTM modelling consistently produced high mean squared error (MSE) and unsatisfactory agreement test results. This study provided two important things more than other previous trials.

Firstly, the study suggested that time-series data alone was not sufficient to predict the real value of BCM-DW. This is understandable considering that patients with different body weights could have similar vital sign patterns during a dialysis session. However, the dry weight of a patient should depend on parameters such as weight, height, or sex, and not just vital signs. We also felt that if the model could predict dry weight accurately without the need of other parameters, its reliability would be questionable.

Secondly, the study found that time-series data had some potential in classifying dialysis sessions based on the gap between pre and post weight. However, this result was limited to certain subpopulations. For example, in configuration A, classifying the subgroup with a high gap was the easiest task. To explore why this was the case, we conducted unsupervised learning of the embedding layer of the LSTM classification model (as shown in Figure 16). We discovered that the embedding layers of this model had the potential to separate these subgroups, as evidenced by the distinct areas of color in the UMAP plot. However, the model failed to do so, possibly due to the lack of some other parameters. Since adverse events were related to the change of vital signs and were not accounted for in the model's learning, further studies are needed to ensure their inclusion in order to maximize the potential of the LSTM model for classification.

Even combining time-series data with non-time-series data, it failed to yield positive outcomes. These findings emphasize that the vital signs parameter alone may not be the sole contributor to body weight. As previously mentioned, they may be potential for predicting gap of weight. However, there were data defects, and many other factors were also at play.

While the study was unable to find a relationship between time-series variables and DW, the findings regarding non-time-series data or general data suggest that it may be worthwhile to further explore this approach. The study was able to develop several models, including gNN, LR, and STACK. Among these models, the STACK model was ultimately selected as it showed the best performance with a small bias from BCM-DW and the narrowest limit of agreement, as also shown in the Folded EDFC.

It should be noted that a more complex neural network model does not always result in better performance. The performance of a model depends on the task at hand, which may be easily captured by a straightforward linear regression or

machine learning model. Additionally, the number of samples required is typically much higher for neural network models.

Although this machine learning model showed generalizability across configuration A and B, it is important to note that there were differences in performance between the two configurations in other models. Configuration B, a model that trained and validated on dialysis sessions in an institute with more dialysis patients, led to easier prediction of dialysis sessions with fewer patients. Therefore, in addition to the number of dialysis samples that affect the model, the number of patients also results in better prediction performance. Thus, while STACK was chosen as the final model for this study, the results may change if applied these models to other datasets with more samples and patients.

5.3 Model's evaluation phase

5.3.1 Graphic analysis

In this section, we focus on assessing each parameter of the Bland-Altman plot of our final model. First and foremost, the mean difference or systemic bias is quite small. This parameter indicates the average difference between the new tool's prediction and the gold standard. From a statistical standpoint, some studies suggest that this number should not exceed 25% of the reference value, as it can result in a number out of limit higher than 5%[53]. However, in clinical practice, this number must be interpreted with caution, as the degree of error in each measurement differs between clinical problems.

As an RCT[26] found that gradual probing of DW can decrease both blood pressure and left ventricular mass index reduction, which are strongly linked to mortality, it is important to have a DW assessment tool with an average difference in mean lower than 1.0 kg. Furthermore, when compared to C-DW performance, the largest trial that compared C-DW and BCM-DW, without testing for agreement, found

that the average euvolemic status by C-DW can be overweighted by BCM-DW by an average gap of 2.0 kg[19]. Our study showed that the final model can accurately predict DW with a small bias from BCM-DW, -0.78 kg in configuration A and +0.02 kg in configuration B. Therefore, the use of ML-DW can provide better prediction performance than C-DW, potentially delivering better clinical outcomes for patients.

The limit of agreement tells us how far the predicted result could deviate from the actual value, and it is usually represented by the upper and lower borders of 1.96 standard deviations for better clinical use. A narrower range of agreement indicates greater confidence in the tool. However, one must keep in mind that this parameter can only be appreciated if the standard tool has consistent results. If the reference tool is good but has volatile results, this will always show a wide range of agreement, even when compared to a good, highly accurate tool.

Since our model was assessed on BCM-DW, which has less variability than C-DW, we can use the result with confidence that it is meaningful. The model showed that the limits of agreement were quite wide, ranging from -3.7 to +2.2 kg in configuration A and -3.5 to +3.5 kg in configuration B. However, compared to studies that use C-DW, which had limits of agreement ranging from 3.79 kg[54] to 7.20 kg[42], our model's limits of agreement were narrower. Therefore, although this ML model cannot completely replace BCM, it has the potential to be a useful tool that provides a narrower limit of agreement than C-DW.

Number out of agreement is another parameter that indicates a tool's consistency compared to the reference tool. It shows the percentage of samples that fall outside the limit of agreement. A lower percentage indicates a higher probability that the new tool can predict results near the mean difference. This is similar to the area under the curve in the folded EDFC, which expects the majority of samples to fall within the prediction zone, with a minority on each tail. Although there is no consensus on the cut point, below 5% is an accepted value in previous studies[42 41] on DW prediction and is also mentioned in some statistics textbooks[53]. In this study,

our final model's results showed borderline numbers of 5.4% (31/570) and 4.9% (29/581) for configurations A and B, respectively. In other words, the predictions from this model had a chance of falling within the acceptable limit as high as 94.6% and 95.1%, which is moderate. This again showed better performance of the model on configuration B, which may be explained by the generalizability of the input data, as already mentioned.

After performing sensitivity analysis and removing the most influential feature towards outliers, in configuration A, a single parameter could not directly result in an outlier. Instead, we suspected that the raw data itself had some variance. Physician decision-making was another factor that influenced the prescription of parameters for each dialysis session but was not included in this study. For instance, if a healthy patient underwent a dialysis process and the physician saw that this patient was doing well, they might try to adjust the net ultrafiltration volume for that session as a trial. Patients might also request their physician to remove more fluid, so they could eat more during the following interdialytic days. Therefore, the parameters that we collected in this study may be affected by the decisions of patients and doctors on a case-by-case basis, which could impact the hidden associations we were seeking to identify.

In configuration B, first sub-analysis showed that parameter removal did decreased number out of agreement in one subgroup but increase in another subgroup instead. So, this model might be at its optimum state and could not be tuning more by this technique. Interestingly, in the other sub-analysis, the removal of Post-dialytic BW and Post-BMI did not result in a significant worsening of the prediction. This could imply that the dependency between this parameter and BCM-DW was not as strong as initially suspected or that the model did a good job of capturing the association based on various parameters, not just a single one. This is a positive result as it suggests that the model is robust and can generalize well to new data.

In summary from graph analysis alone our model was good for low mean difference, which may detect the difference in weight that led to clinical benefit. It showed low number out of limit, but still had wide limit of agreement. Thus, it cannot be a replacement for BCM yet.

5.3.2 Other potentials

To evaluate the model beyond the graph, it is important to consider factors such as time consumption, cost consumption, operator expertise, and generalizability. A new tool may have a lower mean of difference compared to the reference tool, but it may also be more time-consuming or expensive to use, requiring a highly skilled operator. In addition, a tool with a lower mean of difference may not be generalizable to different populations or settings, making it less useful in clinical practice. Therefore, the performance of a new tool must be evaluated not only in terms of its accuracy but also in terms of its practicality and usability in real-world scenarios.

Besides the time required for inputting all the necessary data for the model, the time latency for obtaining the results from the model is around 160 milliseconds, which is less than one minute. This is significantly faster compared to the BCM measurement, which requires the creation of a chip card and takes around 5 minutes from attaching the electrodes to the patient to obtaining the results. Moreover, the patient does not need to be physically present during the test, making it a more convenient option for both patients and physicians. Thus, this tool has the potential to save both time and resources in clinical settings.

Cost is a major limitation of the BCM, especially in developing countries. We have already discussed the cost-effectiveness analysis of the BCM and its impracticality for use in all of our patients, as costs continue to increase. However, a machine learning model can provide a prediction at much lower cost, with a

development process cost of less than five thousand baht, making it a more practical option for our country.

The machine learning model is operator-independent, as it only requires the input of the necessary parameters. In contrast, with the BCM measurement, there are concerns regarding tissue reactance and resistance, which require the assessor to have knowledge of how to maximize electrode conduction, adjust patient position with large body contour to avoid short circuit, and deal with patients with limitations, such as amputees or patients using electrodes not included in the validation study for the BCM. With the ML model, other healthcare workers with less training can use it. Furthermore, if the ML model were trained with a wider range of population using the reference tool, it could be used as a DW assessment tool for a greater number of patients.

Generalizability is a crucial factor when evaluating the effectiveness of a new tool. This cloud-based machine learning model is accessible remotely, eliminating the need for physical equipment or special storage requirements. It is also lightweight, with a size of less than 1MB, making it compatible with mobile devices. This level of accessibility increases its usefulness in a wider range of settings, potentially making it easier for the general population to use the tool.

In addition, despite the lower number of BCM measurements with dialysis sessions per patient in BKIH institute compared to KCMH, the model performed similarly in both Configuration A and B. Therefore, it is possible that the model could be applied to areas with a lower number of BCM measurements per patient. With proper adjustment, this tool could be useful in a wide range of settings.

Furthermore, the ML model has several benefits, including non-invasiveness and high repeatability. The wide limit of agreement may be corrected by repeating the measurement, which is not costly. At this point, we believe that the benefits of the ML model outweigh those of the BCM for various reasons. Although the limit of

agreement may be wide, the mean of agreement is acceptable. While the accuracy may be lower, the model has the potential to be a good replacement if studied with a larger database. Additionally, compared to conventional assessment without any tool, the ML-DW is a complement that can assist physicians in adjusting a patient's DW with more confidence. Its high repeatability could serve as one of the screening tools.

This study has a number of strengths: it was the first machine learning model that aimed to predict BCM-DW. Compared to other models, this one tried to explore the utilization of time-series data in the input variables. It also demonstrated external validation across different institutions. Nevertheless, some limitations remain.

5.4 Limitations and suggestion

Firstly, the retrospective nature of the study made it challenging to demonstrate the effects of all clinical variables on intradialytic changes. Additionally, the lack of data on interventions and consequences during dialysis limited the accuracy of the time-series data in predicting DW. Further studies should focus on collecting more comprehensive data on these factors in a prospective manner.

Another limitation is that the neural network model's performance was poor due to the small number of dialysis sessions in this study. A database from multicenter studies in other countries may be needed to improve this model.

Lastly, it's important to note that the model used in this study may not have included all the relevant parameters that could affect blood pressure and heart rate variation. For example, cardiac function, electrocardiogram, and use of beta-blocker are not incorporated into the model. Therefore, it may be premature to conclude that time-series data is not useful for developing a model for DW prediction. Moreover, collaboration with experts from other subspecialties is necessary to

incorporate a more comprehensive set of features into the model and to determine the feasibility of developing an accurate DW prediction model using time-series data.

5.5 Conclusion

A model from machine learning could assist physicians in DW prediction with a comparable correlation to BCM-DW. However, many factors were needed for its fine-tuning before being implemented in real clinical practice.

This study served as a proof-of-concept that machine learning can be a useful tool for DW prediction but not yet a replacement tool for BCM. With its potential, it may be served as a good screening tool that assist clinical adjustment for DW. This lays the foundation for the development of clinical tests in the future.

REFERENCES

1. A. C. Thailand Renal Replacement Therapy: Year 2020. In: A. L, editor. Thailand Renal Replacement Therapy: The Nephrology Society of Thailand; 2020.

2. National Kidney F. KDOQI Clinical Practice Guideline for Hemodialysis Adequacy: 2015 update. Am J Kidney Dis. 2015;66(5):884-930.

3. Sinha AD, Agarwal R. Can chronic volume overload be recognized and prevented in hemodialysis patients? The pitfalls of the clinical examination in assessing volume status. Semin Dial. 2009;22(5):480-2.

4. Hwang SD, Lee JH, Lee SW, Kim JK, Kim MJ, Song JH. Risk of overhydration and low lean tissue index as measured using a body composition monitor in patients on hemodialysis: a systemic review and meta-analysis. Ren Fail. 2018;40(1):51-9.

5. Flythe JE, Xue H, Lynch KE, Curhan GC, Brunelli SM. Association of mortality risk with various definitions of intradialytic hypotension. J Am Soc Nephrol. 2015;26(3):724-34.

6. Termorshuizen F, Dekker FW, van Manen JG, Korevaar JC, Boeschoten EW, Krediet RT, et al. Relative contribution of residual renal function and different measures of adequacy to survival in hemodialysis patients: an analysis of the Netherlands Cooperative Study on the Adequacy of Dialysis (NECOSAD)-2. J Am Soc Nephrol. 2004;15(4):1061-70.

7. Wizemann V, Schilling M. Dilemma of assessing volume state--the use and the limitations of a clinical score. Nephrol Dial Transplant. 1995;10(11):2114-7.

8. Agarwal R. Hypervolemia is associated with increased mortality among hemodialysis patients. Hypertension. 2010;56(3):512-7.

9. van der Sande FM, Cheriex EC, van Kuijk WH, Leunissen KM. Effect of dialysate calcium concentrations on intradialytic blood pressure course in cardiac-compromised patients. Am J Kidney Dis. 1998;32(1):125-31.

10. Flanigan MJ. Role of sodium in hemodialysis. Kidney Int Suppl. 2000;76:S72-8.

11. Javed F, Savkin AV, Chan GS, Mackie JD, Lovell NH. Recent advances in the monitoring and control of haemodynamic variables during haemodialysis: a review.

Physiol Meas. 2012;33(1):R1-R31.

12. Kraemer M, Rode C, Wizemann V. Detection limit of methods to assess fluid status changes in dialysis patients. Kidney Int. 2006;69(9):1609-20.

13. Flythe JE, Chang TI, Gallagher MP, Lindley E, Madero M, Sarafidis PA, et al. Blood pressure and volume management in dialysis: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int. 2020;97(5):861-76.

14. Thomasset MA. [Bioelectric properties of tissue. Impedance measurement in clinical medicine. Significance of curves obtained]. Lyon Med. 1962;94:107-18.

15. Moissl UM, Wabel P, Chamney PW, Bosaeus I, Levin NW, Bosy-Westphal A, et al. Body fluid volume determination via body composition spectroscopy in health and disease. Physiol Meas. 2006;27(9):921-33.

16. Chamney PW, Kramer M, Rode C, Kleinekofort W, Wizemann V. A new technique for establishing dry weight in hemodialysis patients via whole body bioimpedance. Kidney Int. 2002;61(6):2250-8.

17. Wabel P, Chamney P, Moissl U, Jirka T. Importance of whole-body bioimpedance spectroscopy for the management of fluid balance. Blood Purif. 2009;27(1):75-80.

18. Passauer J, Petrov H, Schleser A, Leicht J, Pucalka K. Evaluation of clinical dry weight assessment in haemodialysis patients using bioimpedance spectroscopy: a cross-sectional study. Nephrol Dial Transplant. 2010;25(2):545-51.

19. Devolder I, Verleysen A, Vijt D, Vanholder R, Van Biesen W. Body composition, hydration, and related parameters in hemodialysis versus peritoneal dialysis patients. Perit Dial Int. 2010;30(2):208-14.

20. Onofriescu M, Hogas S, Voroneanu L, Apetrii M, Nistor I, Kanbay M, et al. Bioimpedance-guided fluid management in maintenance hemodialysis: a pilot randomized controlled trial. Am J Kidney Dis. 2014;64(1):111-8.

21. Liu L, Sun Y, Chen Y, Xu J, Yuan P, Shen Y, et al. The effect of BCM guided dry weight assessment on short-term survival in Chinese hemodialysis patients : Primary results of a randomized trial - BOdy COmposition MOnitor (BOCOMO) study. BMC Nephrol. 2020;21(1):135.

22. Covic A, Ciumanghel AI, Siriopol D, Kanbay M, Dumea R, Gavrilovici C, et al. Value of bioimpedance analysis estimated "dry weight" in maintenance dialysis patients: a systematic review and meta-analysis. Int Urol Nephrol. 2017;49(12):2231-45.

23. Wizemann V, Wabel P, Chamney P, Zaluska W, Moissl U, Rode C, et al. The mortality risk of overhydration in haemodialysis patients. Nephrol Dial Transplant. 2009;24(5):1574-9.

24. Hoppe K, Schwermer K, Klysz P, Radziszewska D, Sawatiuk P, Baum E, et al. Cardiac Troponin T and Hydration Status as Prognostic Markers in Hemodialysis Patients. Blood Purif. 2015;40(2):139-45.

25. Scotland G, Cruickshank M, Jacobsen E, Cooper D, Fraser C, Shimonovich M, et al. Multiple-frequency bioimpedance devices for fluid management in people with chronic kidney disease receiving dialysis: a systematic review and economic evaluation. Health Technol Assess. 2018;22(1):1-138.

26. Agarwal R, Bouldin JM, Light RP, Garg A. Probing dry-weight improves left ventricular mass index. Am J Nephrol. 2011;33(4):373-80.

27. Hur E, Usta M, Toz H, Asci G, Wabel P, Kahvecioglu S, et al. Effect of fluid management guided by bioimpedance spectroscopy on cardiovascular parameters in hemodialysis patients: a randomized controlled trial. Am J Kidney Dis. 2013;61(6):957-65.

28. Huan-Sheng C, Yeong-Chang C, Ming-Hsing H, Fan-Lieh T, Chu-Cheng L, Tsai-Kun W, et al. Application of bioimpedance spectroscopy in Asian dialysis patients (ABISAD-III): a randomized controlled trial for clinical outcomes. Int Urol Nephrol. 2016;48(11):1897-909.

29. Carter M, Morris AT, Zhu F, Zaluska W, Levin NW. Effect of body mass index (BMI) on estimation of extracellular volume (ECV) in hemodialysis (HD) patients using segmental and whole body bioimpedance analysis. Physiol Meas. 2005;26(2):S93-9.

30. Kyle UG, Bosaeus I, De Lorenzo AD, Deurenberg P, Elia M, Manuel Gomez J, et al. Bioelectrical impedance analysis-part II: utilization in clinical practice. Clin Nutr. 2004;23(6):1430-53.

31. Furstenberg A, Davenport A. Comparison of multifrequency bioelectrical impedance analysis and dual-energy X-ray absorptiometry assessments in outpatient

hemodialysis patients. Am J Kidney Dis. 2011;57(1):123-9.

32. Scharfetter H, Wirnsberger GH, Holzer H, Hutten H. Influence of ionic shifts during dialysis on volume estimations with multifrequency impedance analysis. Med Biol Eng Comput. 1997;35(2):96-102.

33. Di Iorio BR, Scalfi L, Terracciano V, Bellizzi V. A systematic evaluation of bioelectrical impedance measurement after hemodialysis session. Kidney Int. 2004;65(6):2435-40.

34. Jacobsen E, Cruickshank M, Cooper D, Marks A, Brazzelli M, Scotland G. Cost-effectiveness and value of information analysis of multiple frequency bioimpedance devices for fluid management in people with chronic kidney disease having dialysis. Cost Eff Resour Alloc. 2021;19(1):24.

35. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015;115:211-52.

36. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural computation. 2006;18(7):1527-54.

37. Niel O, Bastard P, Boussard C, Hogan J, Kwon T, Deschenes G. Artificial intelligence outperforms experienced nephrologists to assess dry weight in pediatric patients on chronic hemodialysis. Pediatr Nephrol. 2018;33(10):1799-803.

38. Kim HR, Bae HJ, Jeon JW, Ham YR, Na KR, Lee KW, et al. A novel approach to dry weight adjustments for dialysis patients using machine learning. PLoS One. 2021;16(4):e0250467.

39. Chiu JS, Chong CF, Lin YF, Wu CC, Wang YF, Li YC. Applying an artificial neural network to predict total body water in hemodialysis patients. Am J Nephrol. 2005;25(5):507-13.

40. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb). 2015;25(2):141-51.

41. Guo X, Zhou W, Lu Q, Du A, Cai Y, Ding Y. Assessing Dry Weight of Hemodialysis Patients via Sparse Laplacian Regularized RVFL Neural Network with L(2,1)-Norm. Biomed Res Int. 2021;2021:6627650.

42. Guo X, Zhou W, Yu Y, Cai Y, Zhang Y, Du A, et al. Multiple Laplacian Regularized

RBF Neural Network for Assessing Dry Weight of Patients With End-Stage Renal Disease. Front Physiol. 2021;12:790086.

43. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. nature. 1986;323(6088):533-6.

44. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735-80.

45. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.

46. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). 2005;67(2):301-20.

47. Schölkopf B, Smola AJ, Bach F. Learning with kernels: support vector machines, regularization, optimization, and beyond: MIT press; 2002.

48. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.

49. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems. 2017;30.

50. Wolpert DH. Stacked generalization. Neural networks. 1992;5(2):241-59.

51. Dwarampudi M, Reddy N. Effects of padding on LSTMs and CNNs. arXiv preprint arXiv:190307288. 2019.

52. Monti KL. Folded empirical distribution function curves—mountain plots. The American Statistician. 1995;49(4):342-5.

53. Fraser CG. Biological variation: from principles to practice: Amer. Assoc. for Clinical Chemistry; 2001. 52-5 p.

54. Jian Y, Li X, Cheng X, Chen Y, Liu L, Tao Z, et al. Comparison of bioimpedance and clinical methods for dry weight prediction in maintenance hemodialysis patients. Blood purification. 2014;37(3):214-20.

VITA

NAME	Nataphut Boonvisuth
DATE OF BIRTH	7 September 1990
PLACE OF BIRTH	Khon Kaen
HOME ADDRESS	318/25 Noble Tara Village, Rattanathibet Road, Bang Kraso, Muang, Nonthaburi, 11000