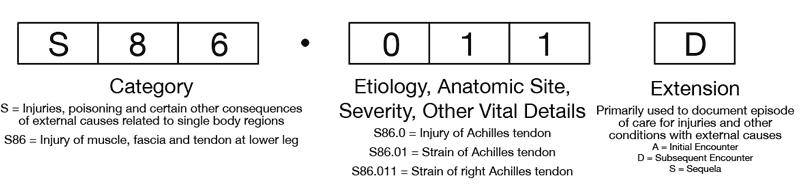
**Data Sources**

For this thesis, datasets from the Medical Information Mart for Intensive Care version 1.4 (MIMIC-III v1.4) database, which is available on PhysioNet Clinical Databases [https://mimic.physionet.org](https://mimic.physionet.org/) was used. MIMIC-III is a large, open-access, well maintained database with a collection of deidentified health-related data associated with over 58,000 hospital admissions for 38,645 adults and 7,875 neonates, who was admitted to critical care units of the Beth Israel Deaconess Medical Center (Boston, MA, USA) between the period of June 2001 and October 2012 and is hosted by the Laboratory for Computational Physiology (LCP) at the Massachusetts Institute of Technology (MIT). [cite physionet] MIMIC-III contains clinical, physiological and mortality data was aggregated from ICU information systems, digital hospital archives, bedside monitors and Social Security Administration Death Master Files. MIMIC-III includes health-related information such as demographics, vital sign measurements, laboratory test results, procedures, medications, caregiver notes, imaging reports, and in and out of hospital mortality registry. Data were collected continuously with non-invasive procedures during routine clinical care and data collection and was blinded from patient and staff, minimising interference with medical routine. Private health information was also de-identified to protect patient privacy. [cite 2 https://link.springer.com/chapter/10.1007/978-3-319-43742-2\_5]

Access and use of the MIMIC-III database for this thesis has been authorised with a restricted-access to PhysioNet clinical databases and is bound by data use agreement, approved on 9th April 2019.

**International Classification of Diseases ICD 9**

International Statistical Classification of Diseases and Related Health Problems (ICD) is the international standard for clinical diagnosis. ICD defines diseases, disorders, injuries and other related health conditions in a hierarchical fashion, this allows for storage and access of health-related information internationally, in other words ICD is the standard for clinical diagnosis. [https://www.who.int/classifications/icd/en/] Though there are subtle variation to the standard as some country adapt different versions. For example, Clinical Modification (ICD-9-CM) is an adaption used in assigning diagnostic in the United States[https://www.cdc.gov/nchs/icd/icd9cm.htm], and in New Zealand, the Australian Modification (ICD-10-AM) were used. The version of ICD used for medical diagnosis in MIMIC dataset is ICD-9, the codes are assigned at the end of the patient’s stay and are used by the hospital to bill for care provided. [ref mimic3]



https://www.webpt.com/blog/post/understanding-icd-10-code-structure

The first three characters of an ICD codes designate the category of the diagnosis, followed by supplementary information of etiology (i.e., the cause of condition), anatomic site, severity, or other vital clinical details after a full stop, and may contain an extension used to document episodes of care, this was implemented for ICD 10. For some instances, there may be multiple codes entries for a single condition, like in the case of an Achilles tendon strain, secondary external cause code may be provided along with the primary diagnosis.

Insert table

ICD 9 categories are coded from 001 to 999, with two extra supplementary class of categories starting with V and E. For example, 001 is the code for Cholera and E880 is the code for Accidental fall on or from stairs or steps. Several categories are aggregated to form a larger category, and there are 2 aggregations which made ICD a three-level hierarchical structure. Following Cholera at level3, it is aggregated into Intestinal Infectious Diseases for categories from 001 to 009 at level 2, and Infectious and Parasitic Diseases for the categories from 001 to 139 at level 1.

ICD helps to provide a picture of the health situation of the general populations, this information is useful in health care management, and allocation of resources, and other health related decision-making process, such as billing of services.

**Study Population**

<https://www.r-bloggers.com/flow-charts-in-r/> **consort diagram**

Study inclusion criteria were ICU admission of patients of all age, including repeated admission by same patient, with available documentation of International Classification of Diseases (ICD-9), only the primary diagnosis were considered.

16 years and older and available documentation of height and weight, as well laboratory test results at baseline and at ICU admission.

We excluded data from subsequent admissions if patients were admitted to ICU more than once. We defined baseline laboratory values as the mean laboratory result of all readings available between 3 days before to 1 year prior to ICU admission. ICU values were defined as the most abnormal laboratory result in the first 24 hours of ICU admission, similar to the analysis in the calculation of the SAPS-II and SOFA scores (5, 6). Obesity was determined according to World Health Organization classification. The height measured during the hospital admission and the average of weights measured 24 hours before and 24 hours after the ICU admission were used. Only obese (body mass index [BMI], ≥ 30) and normal weight patients (BMI, ≥ 18.5 and < 25) were included as comparison groups to maximize the difference between study groups.

The final dataset used contain……

**Study Variables**

Merging with three datasets, including the admissions table, the diagnoses\_icd table and the d\_icd\_diagnoses table,

the following baseline patient level characteristics were collected: age, gender, ethnicity, marital status, insurance coverage, and comorbidities as defined by Elixhauser et al (14) combined in a composite score by van Walraven et al (15), here after referred to as the comorbidity index.

Smoking status was identified using Natural Language Processing searches for history of active smoking in the provider notes. Hospital characteristics, procedures in the first 24 hours of the ICU admission, as well as SAPS-II and SOFA score on ICU admission were also included (5, 6).

The exposure variable was BMI status, comparing obese to normal weight individuals, and the primary outcome was the deviation in laboratory results between that measured at baseline and during ICU admission.

We selected laboratory results that were used in the SAPS-II or SOFA scores in our analyses, except for bilirubin, which was not included due to a significant fraction of missing baseline data.

**Statistical Analyses**

We used quantile-quantile normal plots to assess the appropriateness of assuming normality. Continuous variables were summarized using the mean and sd while those with a nonnormal distribution were summarized with the median and interquartile range. For the continuous variables, mean values were compared using two-sample t tests, and median values were assessed using the Mann-Whitney test. Tests for association between categorical variables and BMI status were assessed using a chi-square test. Absolute values at baseline as well as deviations from baseline were compared between normal weight and obese individuals. The differences in deviation from baseline between

both groups were compared using multivariable linear regression adjusted for age, gender, comorbidity index (15), SAPS-II score (5) or SOFA score (6), and type of ICU, and the relevant baseline laboratory result. A full model comprising the BMI status and all the covariates was initially fit and subjected to stepwise backward elimination retaining BMI status in the model, until a final model was obtained with only statistically significant variables. Statistical significance was assessed at the 0.05 level. For variables violating the modeling assumptions of linear regression models, the logarithm (base 10) of the baseline and ICU laboratory results were calculated and the regression analysis performed on the log transformed values. We also assessed the effect on hospital mortality of any statistically significant deviations found comparing the obese and normal weight groups using logistic regression. A null (baseline) model was fit composed of SAPS-II, SOFA, age, and the ICU values of the laboratory tests, which were found to be statistically significant when comparing the deviations from baseline between obese and normal weight subjects. A model fit using all variables in the null model, in addition to any laboratory deviation variables found to have a statistically significant difference between the normal weight and obese patients, was compared with the null model using a likelihood ratio test. Information about the number of missing laboratory values is provided in detail in Supplemental Table 1 (Supplemental Digital Content 1, http://links.lww.com/CCM/D72). All

SAS version (ref) was used during the data cleaning phase, and analyses were performed with R version 3.3.2 (R Foundation for Statistical Computing, Vienna, Austria). Extracts of coding is available in the appendix and We have made all our data extraction and modeling queries and codes available on GitHub: