Analysis plan

2019-02-04
Martin Henriksson
073-070 73 25
martin.henriksson@stud.ki.se

Research question

How does the transfer of clinical prediction models for early trauma care between different contexts within a single health care system affect mistriage rates?

Data set

Data from the Swedish Trauma registry SweTrau will be used for the statistical analysis. SweTrau is a nationally encompassing registry in which 95,5% (52 of 55) of Swedish hospitals record trauma cases. Currently the registry contains 55 000 cases. Patient data is recorded by hospital personnel. Inclusion criteria are: Traumatic event with subsequent activation of hospital trauma protocol, admitted patients with NISS > 15 and patients transferred to the hospital within 7 days of traumatic event with NISS > 15. SweTrau excludes patients if the only traumatic event is chronic subdural hematoma or if hospital trauma protocol is activated without traumatic event (1).

Inclusion criteria

The inclusion criteria are patients registered in SweTrau and age over 15. This age was decided as the study aims to study adult trauma and not paediatric trauma which differs in triage and initial care (2). Patient age will be obtained from the SweTrau registry.

Variables

Model Predictors

Our clinical prediction models will include the predictors systolic blood pressure (SBP), respiratory rate (RR) and Glasgow Coma Scale (GCS) on arrival to hospital. The rationale for including these three predictors is that they are part of many established clinical prediction models for early trauma care, such as the Revised Trauma Score(3). SBP and RR will be modelled using restricted cubic splines with four knots placed at equally spaced percentiles and GCS as a continuous linear term. When describing the participant characteristics all quantitative variables will be presented as continuous. ISS will also be presented as dichotomous using ISS > 15 as the cutoff.

Model outcome

The outcome that will be used to develop the clinical prediction models is all cause mortality within 30 days of the trauma.

Participant characteristics

To describe the patient cohort we will report age, sex, American Society of Anaesthesiologists physical status classification system (ASA), Injury Severity Score (ISS) and New Injury Severity Score (NISS).

Study outcome

We will use ISS > 15 as the gold standard to define trauma severity as major trauma, and hence patients with ISS ≤ 15 will be considered minor trauma (4). We define overtriage as the event when a clinical prediction model classifies a patient with ISS ≤ 15 as major trauma, and undertriage as the event when a clinical prediction model classifies a patient with ISS > 15 as minor trauma. We define the overtriage rate as the number of overtriaged patients divided by all patients. We define the undertriage rate as the number of undertriaged patients divided by all patients. The mistriage rate is defined as the sum of the over- and undertriage rates.

Statistical methods and software

Initial data management

The programming language R will be used for all analyses (5). The SweTrau data set will be imported, and a data frame created. From the complete SweTrau data set only the columns of variables of interest will be selected: Predictors; SBP upon arrival to hospital, RR upon arrival to hospital and GCS upon arrival to hospital. Outcome; 30-day mortality (Survival status). Covariates; Age, sex, ASA-classification, ISS, NISS, date of trauma, clinic number (Klinik nummer) and hospital code (Sjukhuskod). See table 1.

Variable	Definition	Abbreviated field name	Missing data term
SBP	First recorded SBP upon arrival in the ED / hospital.	ed_sbp_value	NA
RR	First recorded RR upon arrival in the ED / hospital.	ed_rr_value	NA
GCS	First recorded GCS score upon arrival in the ED / hospital.	ed_gcs_sum	999 or NA
30-day mortality (Survival status)	Dead or alive 30 days after trauma	res_survival	999 or NA
Age	The patient's age at the time of injury.	pt_age_yrs	NA
Sex	The patient's gender	pt_gender	999 or NA
ASA	The co--morbidity existing before the incident.	pt_asa_preinjury	999 or NA
ISS	Injury Severity Score	ISS	?
NISS	New Injury Severity Score	NISS	?
Date of trauma	Date of trauma	DateTime_Of_Trauma	NA
Clinic number	?	kli_KlinikNr	?
Hospital code	?	Sjukhuskod	?

Table 1: Variables of interest. Variables used for the creation, validation and comparison of the clinical prediction models.

Data cleaning

Predictor variables

By using the function rcspline.eval(x, nk = 4) as implemented in the R package Hmisc, SBP and RR will be converted to restricted cubic spindles with four knots placed at equally spaced percentiles. Knot locations from the development samples will be stored so that the same knot locations can be used in the validation samples. No changes will be made to GCS, it will be kept as a continuous variable.

Abnormal entries

GCS with a value of 99 (Intubated) will be replaced with 3, this is done to enable us to better include these patients in the prediction model. It is reasonable to assume the intubated patient has a GCS of 3. Other values that appear to differ from the data structure will also be identified and dealt with in an appropriate manner.

Missing data

Different variables use different expressions for missing data, see table 1. To handle missing data properly, all these will be replaced with NA. We will use multiple imputation using chained equations, as implemented in the R package mice, to handle missing data (6). The missing data is assumed to be missing at random (MAR). The number of imputations to be created for each data set will be equal to the percentage of missing data in that data set. Quantitative variables will be imputed using predictive mean matching and qualitative variables will be imputed using logistic regression. SBP and RR will be transformed as restricted cubic splines before imputation and imputed as just another variable. We will present the combined results as the median point estimate along with the minimum value of the lower and maximum value of the upper 95% CI bounds across imputations. This combined CI will henceforth be referred to as CI_{MI}. This process will be repeated in all data sets prior to data analysis.

Data sets and samples

The complete SweTrau cohort will be split into four sets of data. High and low volume centres, metropolitan and non-metropolitan centres, multi and single centre data and individual centres. The process of identifying and separating these datasets is outlined below.

High and low volume centres

Based on number of patients, two samples will be derived from this data set. High volume centres will be those with in the top quartile of number of patients received. The rest will be low volume centres. To achieve this split the hospitals with the top quartile number of patients registered will be identified, this information will then be cross-referenced with the clinic number and hospital code in SweTrau. The cases with these clinic numbers and hospital codes will constitute the “High volume centre sample”, the rest will be the “Low volume centre sample”.

Metropolitan and non-metropolitan centres

This data set will also be split into two samples. The metropolitan sample will consist of greater Stockholm, greater Gothenburg and greater Malmö, as defined by statistics Sweden. The other sample will be patients from non-metropolitan areas. Once again, the clinic number and hospital code in SweTrau will be used to identify cases belonging to hospitals in the stated regions, this will be the “Metropolitan sample”, and the rest will be the “Non-metropolitan sample”.

Multi and single centre data

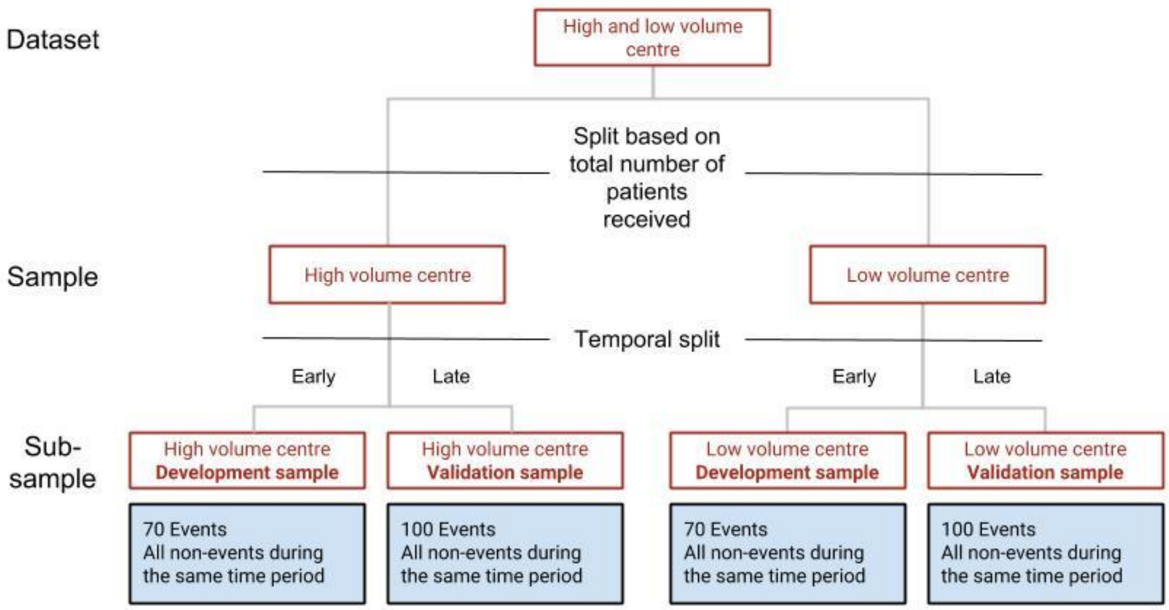
In this data set multiple samples will be created. Each centre with large enough sample size to develop and validate a model will constitute their own sample. The multi-centre sample will consist of the combined data from all single centre samples. By using the clinic number and hospital code, hospitals with at least 170 events (events being patients who died within 30 days of the trauma) will be identified. Cases belonging to each of these hospitals will constitute their own “Single centre sample”. A combination of all these “Single centre samples” will be used to create the “Multi centre sample”.

Individual centres

This data set will also be split into multiple samples. Each centre with large enough sample size to develop and validate a model will constitute its own sample. By using the same method as in the Multi and single centre data, hospitals with at least 170 events will be identified. Cases belonging to each of these hospitals will constitute their own “Individual centre sample”.

Subsamples

Each set of data will thus include at least two samples. The samples will then further be split into two subsamples using a temporal split based on the date of traumatic event. This is achieved by using the *order(x)* function of R to sort by date of trauma. The earlier subsample will be the development sample, and the later subsample the validation sample. The development sample will contain 70 events and all non-events during the same time. The rationale for including 70 events is that we need at least 10 events per free parameter in the logistic regression to obtain stable coefficient estimates (9). The validation sample will contain 100 events and at least 100 non-events (10). See figure 1 for example.



Dataset
High and low volume centre
Split based on total number of patients received
Sample
High volume centre
Low volume centre
Temporal split
Early
Late
Early
Late
Sub-sample
High volume centre Development sample
High volume centre Validation sample
Low volume centre Development sample
Low volume centre Validation sample
70 Events All non-events during the same time period
100 Events All non-events during the same time period
70 Events All non-events during the same time period
100 Events All non-events during the same time period

Figure 2: High- and low volume centre data set. Initial split based on number of patients. Temporal split made using date of traumatic event.

The total number of events and non-events per data set will therefore be at least 680. The minimum sample size of development and validation samples will be 140 and 200 respectively. We will perform analyses only on data sets for which all samples include at least the minimum number of patients.

Data analysis

We will perform the analyses in the sequence of model development, model validation and finally model comparison. These steps will be repeated in each data set. Below we use the transfer of a model from a high-volume sample to a low volume sample as an example to describe the complete procedure.

Model development

In the model development step a clinical prediction model will be developed in the high-volume centre development sample. The model will be developed using logistic regression as implemented in the R function glm. The dependent variable will be all cause mortality within 30 days of trauma and independent variables will be SBP, RR, and GCS modelled as previously described. To avoid overfitting the model we will use a bootstrap procedure to estimate a linear shrinkage factor that we will apply to the model coefficients (7). The shrunk model will then be used to estimate the probability of all cause 30-day mortality in the development sample. We will then do a gridsearch across estimated probabilities in the development sample to identify the cut-off that optimises overtriage keeping undertriage at less than 5% (8). This cut-off will then be used to classify patients as major or minor trauma.

Model validation

In the model validation step, the model performance will be assessed in the high-volume centre validation sample and in the low volume centre validation sample. First the model will be used to estimate the probability of all cause 30-day mortality in each sample. Then the probability cutoff identified in the development sample will be applied to the validation samples, patients will be classified as major or minor trauma, and model performance is estimated.

Model comparison

Finally, in the model comparison step, the difference in model performance between the high and low volume centre validation samples will be calculated. We will use an empirical bootstrap to estimate 95% confidence intervals (CI) around performance and differences in performance estimates. Both bootstrap procedures used will use 1000 bootstrap samples drawn with replacement of the same size as the original samples.

Performance measures

Model performance will be assessed in terms of over-, under-, and mistriage rates as defined above.

Stepwise description of analysis

1. Import study data. Load R package: mice and Hmisc.
2. Create study sample by only keeping relevant variables
3. Data
    a. Convert SBP and RR using restricted cubic splines
    b. Replace GCS 99 with 3. Replace all variants of entries denoting missing data with NA.
    c. Manage missing data with multiple imputation using chained equations, via the R package mice. Use predictive mean matching for quantitative variables and logistic regression for qualitative variables. Present combined results as specified.

4. Summarize cleaned data frame and save results. This data frame will from now be known as cleaned data
5. Create data sets and samples
    a. High and low volume data set
        i. Split data using clinic number and hospital code. The high volume sample will be those belonging to the quartile of centres with the largest number of patients registered.
        ii. Split high volume sample using date of trauma, specify development sample and validation sample
        iii. Split low volume sample using date of trauma, specify development sample and validation sample
    b. Metropolitan and non-metropolitan data set
        i. Split data using clinic number and hospital code. Use kli_KlinikNr and Sjukhuskod to identify hospitals in Greater Stockholm, Greater Gothenburg and Greater Malmö.
        ii. Split metropolitan sample using date of trauma, specify development sample and validation sample
        iii. Split non-metropolitan sample using date of trauma, specify development sample and validation sample
    c. Multi and single centre data set
        i. By using clinic number and hospital code, identify each centre with large enough sample size. Make each of these their own sample
        ii. Combine all single centre samples into a multi centre sample
        iii. Split each single centre sample using date of trauma, specify development sample and validation sample
        iv. Split multi centre sample using date of trauma, specify development sample and validation sample
    d. Individual centre data set
        i. By using clinic number and hospital code, identify each centre with large enough sample size. Make each of these their own sample
        ii. Split each individual centre sample using date of trauma, specify development sample and validation sample
6. Make sure all data sets and samples are correctly named and organized. Create backup copies of all progress both locally and using secure online storage

7. Begin data analysis
8. Model development
    a. In the development sample use the glm function to create a logistic regression model that includes SBP, RR and GCS as the independent variables and 30-day-mortality as the dependant variable
    b. A bootstrap procedure will be used to estimate a linear shrinkage factor which will then be applied to the model coefficients
    c. By using the model, 30-day-mortality will be estimated in the development sample
    d. A gridsearch will be performed across the estimated probabilities to identify the cut-off that optimises overtriage keeping undertriage at less than 5%
    e. Using this cut-off patients in the development sample will be identified as major or minor trauma
9. Model validation
    a. Use the model developed in step 8 to estimate the probability of 30-day-mortality in the validation sample (all validation samples in the data set)
    b. The cut-off identified in the development sample is then applied to the validation sample and patients classified as major or minor trauma
    c. Model performance in terms of over and undertriage is saved
10. Model comparison
    a. The difference in model performance between each validation sample is calculated and saved
11. Steps 8 to 10 will be repeated for all data sets
12. Step 11 is repeated in 1000 bootstrap samples drawn with replacement to estimate 95% CI around performance and performance differences
13. Summarize and save performance and comparison data
14. Assess the models used in each data set. Determine mistriage, overtriage and undertriage as well as the rates of each.

References

1. SweTrau. MANUAL SVENSKA TRAUMA REGISTRET [Internet]. Stockholm: Svenska Traumaregistret; 2018 [cited 2019 2019-01-25]. Available from: http://rcsyd.se/swetrau/wp-content/uploads/sites/10/2019/01/SweTrau-Manual-2018.pdf.

2. McFadyen JG, Ramaiah R, Bhananker SM. Initial assessment and management of pediatric trauma patients. Int J Crit Illn Inj Sci. 2012;2(3):121-7.

3. Champion HR, Sacco WJ, Copes WS, Gann DS, Gennarelli TA, Flanagan ME. A revision of the Trauma Score. J Trauma. 1989;29(5):623-9.

4. Palmer C. Major trauma and the injury severity score--where should we set the bar? Annu Proc Assoc Adv Automot Med. 2007;51:13-29.

5. R-foundation. The R Project for Statistical Computing [Internet]. 2018 [cited 2019 2019-01-25]. Available from: https://www.r-project.org/.

6. van Buuren S, Groothuis-Oudshoorn C. MICE: Multivariate Imputation by Chained Equations in R2011.

7. Steyerberg EW, Eijkemans MJC, Habbema JDF. Application of Shrinkage Techniques in Logistic Regression Analysis: A Case Study. Statistica Neerlandica. 2001;55(1):76-88.

8. Rotondo MF, Cribari C, Smith RS. Resources for the optimal care of the injured patient. Chicago: American College of Surgeons; 2014.

9. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373-9.

10. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58(5):475-83.