# Term Project #1 Fall, 2021

## Due Date: 11:59PM 12/12/2021

**Data mining in PJI research: periprosthetic joint infection early diagnosis for total joint replacement**

**Background**

Total knee/hip joint replacement (total knee/hip arthroplasty) is performed to restore function and relieve pain in patients with severely damaged knees. The surgery involves replacement of both the medial and lateral femorotibial joints and the patellofemoral joint. Although total joint replacement is an effective treatment, postoperative complications include blood clots, infection, and loosening or malalignment of the prosthetic component. Periprosthetic joint infection (PJI) is a serious complication occurring in 1% to 2% of primary arthroplasties, which is associated with high morbidity and need for complex interdisciplinary treatment strategies.

## Objective

This Project aims to encourage the development of algorithms for detecting the periprosthetic joint infection (PJI) from medical records. An accurate early PJI diagnosis can help doctors proceed with further necessary and appropriate treatments.

## Data Description

We have posted 52159 of the joint replacement surgical patient records and labels as public training sets and kept 980 records as private test sets. For more details about the data for this project, please see Table 1.

## Evaluation

Different measures, as defined in Table 2, can be used to evaluate the predictive performances. You can use the *appropriate* measures to evaluate the performances of your models.

## Mining tool

C4.5 developed by Quinlan, as posted on e3.

C4.5 was originally developed on UNIX. To make it run on different operating systems (e.g. Windows), you may revise the C4.5 source code a bit, but only enough that you can compile and run it on your platform.

You are NOT allowed to modify the core of the source.

## Evaluation and What to Turn in

Your project will be evaluated on 2 parts:

- Predictive performances.
  Your predictions will be evaluated based on various *appropriate* measures.
- A detailed report that clearly describes your data pre-preprocess and any other necessary data-centric procedure.
  Procedures may involve, but not limited to: (1) Data Cleaning; (2) Data Transformation; (3) Data Reduction; (4) Data imputation, etc.
  For this part, you must (1) discuss possible problems, and (2) explain how to deal with them?
  Some hints:
    - ◆ Provide summary statistics, box plot, and histogram… Detect if there exist any outliers or anomalies in this dataset.
- You need to turn in your final training dataset and test dataset in Excel sheet format.
- You need to turn in your a print-out (in text file) of your final trained classification tree.

PS. Term Project #2 is coming soon.

**Table 1.** Data Description

| Item | description | | Item | description |
|---|---|---|---|---|
| Outcome | infected or non-infected | | Cemented | procedure |
| Non_commercial_ALBC | Noncommercial antibiotic-loaded bone cement (procedure) | | Commercial_ALBC | Commercial antibiotic-loaded bone cement (procedure) |
| Age | Age | | Sex | Sex |
| LOS | length of stays | | CBC_WBC | routine blood test |
| Joint | Categories of Joint | | CBC_RBC | routine blood test |
| Drain | drainage (procedure) | | CBC_HG | routine blood test |
| cci_index | Charlson Comorbidity Index | | CBC_HT | routine blood test |
| elx_index | Elixhauser Comorbidity Index | | CBC_MCV | routine blood test |
| Blood_trans | Blood transfusion | | CBC_MCH | routine blood test |
| OP_time_minute | Operation time (minutes) | | CBC_MCHC | routine blood test |
| OP_time_hour | Operation time (hours) | | CBC_RDW | routine blood test |
| ASA | ASA code | | CBC_Platelet | routine blood test |
| Congestive Heart Failure | heart failure | | CBC_RDWCV | routine blood test |
| Cardiac Arrhythmia | irregular heartbeat | | BUN | routine blood test |
| Valvular Disease | Heart valve disease | | Crea | routine blood test |
| Heart disease | such as coronary heart disease, heart attack, congestive heart failure, and congenital heart disease. | | GOT | routine blood test |
| Pulmonary Circulation Disorders | Pulmonary vascular disease | | GPT | routine blood test |
| Peripheral Vascular Disorders | Peripheral vascular disease | | ALB | routine blood test |
| Hypertension Uncomplicated | Hypertension Uncomplicated | | Na | routine blood test |
| Paralysis | Paralysis | | K | routine blood test |
| Other Neurological Disorders | Neurological Disorders | | UA | Uric Acid |
| Chronic Pulmonary Disease | Chronic obstructive pulmonary disease | | Diagnosis | {1, 2, 3, 4, 5} |

| Item | description | | Item | description |
|---|---|---|---|---|
| Lung disease | Lung disease | | Metastatic Cancer | Metastatic Cancer |
| Diabetes | Diabetes mellitus | | Solid Tumor without Metastasis | Solid Tumor without Metastasis |
| Hypothyroidism | Hypothyroidism | | Cancer history | Cancer history |
| Renal Failure | Renal Failure | | Rheumatoid Arthritis/collagen | Rheumatoid Arthritis/collagen |
| Liver Disease | Liver Disease | | Coagulopathy | Coagulopathy |
| Peptic Ulcer Disease excluding bleeding | Peptic Ulcer Disease excluding bleeding | | Obesity | Obesity |
| AIDS/HIV | Acquired Immune Deficiency Syndrome | | Weight Loss | Weight Loss |
| Lymphoma | Lymphoma | | Fluid and Electrolyte Disorders | Fluid and Electrolyte Disorders |

| | | | | |
|---|---|---|---|---|
| Psyciatric disorder | Psyciatric disorder | | Blood Loss Anemia | Blood Loss Anemia |
| Anemia | Anemia | | Deficiency Anemia | Deficiency Anemia |
| Alcohol Abuse | Alcohol Abuse | | Drug Abuse | Drug Abuse |
| Psychoses | Psychoses | | Depression | Depression |

**Table 2.** Evaluation Metrics

| True Positive，TP<br>(Infected) | True Negative，TN<br>(Non-infected) | False Positive，FP | False negative，FN |
|---|---|---|---|

| Indicator | Definition |
|---|---|
| Accuracy，ACC | $\frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN})}$ |
| F1-Score (Based on Death)<br>(Harmonic mean of Precision and Sensitivity) | $\frac{(2*TP)}{(2*TP+FP+FN)}$ |
| Positive Predictive Value, PPV (Precision) | $\frac{(TP)}{(TP+FP)}$ |
| True positive rate，Recall，Sensitivity | $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$ |
| Matthew's Correlation Coefficient，MCC | $\frac{(\mathrm{TP}*TN)-(FP*FN)}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$ |
| Area Under the ROC curve, AUC | Area under the ROCcurve |