Imperial College London

**Final Year Project 2021**

Thursday, 24 June 2021

# Analysis and Implementation of Data Imputation Techniques for Laboratory Data

Supervisor: Dr Pantelis Georgiou
Co-supervisor: Dr Bernard Hernandez Perez
Second marker: Dr Timothy Constandinou

Agrim Manchanda (CID: 01333674)

Imperial College London

# Overview of presentation

- Introduction and Motivation
- Background Materials
- Project Aims and Deliverables
- Overview of Design & Implementation
- Experiments & Results
- Conclusion, Achievements and Future Work

Imperial College London

# Introduction and Motivation

- Missing laboratory data is an ***ever-present*** challenge in clinical domain.
- Clinical Decision Support Systems (CDSSs) rely on ***completeness*** for ***accurate*** predictions.
- Simple techniques exist but are ***inadequate***.
- Scope to investigate **better** data imputation techniques suitable for laboratory data.

| *Patient* | *Date* | *Lab Code* | | | | |
|---|---|---|---|---|---|---|
| | | *EOS* | *MONO* | *BASO* | *....* | *NRBCA* |
| 1 | 23/03/20 | 0.40 | *NaN* | 0.10 | .... | 0.00 |
| 2 | 24/03/20 | *NaN* | 0.50 | 0.10 | .... | *NaN* |
| 2 | 23/03/20 | *0.20* | *NaN* | 0.10 | .... | 0.00 |
| 3 | 23/03/20 | 0.40 | 0.50 | *NaN* | .... | 0.00 |
| 4 | 23/03/20 | 0.10 | *NaN* | 0.10 | .... | *NaN* |
| 5 | 23/03/20 | NaN | 0.40 | 0.10 | .... | 0.00 |

| *Patient* | *Date* | *Lab Code* | | | | |
|---|---|---|---|---|---|---|
| | | *EOS* | *MONO* | *BASO* | *....* | *NRBCA* |
| 1 | 23/03/20 | 0.40 | *0.47* | 0.10 | .... | 0.00 |
| 2 | 24/03/20 | *0.28* | 0.50 | 0.10 | .... | *0.00* |
| 2 | 23/03/20 | *0.20* | *0.47* | 0.10 | .... | 0.00 |
| 3 | 23/03/20 | 0.40 | 0.50 | *0.10* | .... | 0.00 |
| 4 | 23/03/20 | 0.10 | *0.47* | 0.10 | .... | *0.00* |
| 5 | 23/03/20 | 0.28 | 0.40 | 0.10 | .... | 0.00 |

Source: [1]: Bernard Hernandez Perez. *Data-driven web-based intelligent decision support system for infection management at point of care*. PhD thesis, Imperial College London, 2 2019.

Imperial College London

## Background Material (1)

- Types of missing data:
  - Missing at Random (MAR)
  - Missing Completely at Random (MCAR)
  - Missing Not At Random (MNAR)

- Current landscape:
  - Imputation **depends** on feature types and **nature** of missing data.
  - Separate studies **validated** Machine Learning (ML) methods.
  - Increasing **potential** for use of Bayesian Networks (BN).

Imperial College London

# Background Material (2)

## *Machine Learning*

Keys:
Dependent Variable
Independent Variables

| A | B | C |
|---|---|---|
| … | … | … |
| … | … | … |
| … | … | … |

| A | B | C |
|---|---|---|
| … | … | … |
| … | … | … |
| … | … | … |

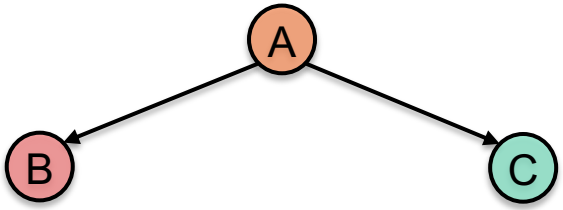
| A | B | C |
|---|---|---|
| … | … | … |
| … | … | … |
| … | … | … |

| A | B | C |
|---|---|---|
| … | … | … |
| … | … | … |
| … | … | … |

**Model Learning → Evaluation**

## *Bayesian Networks*

| Yes | No |
|---|---|
| … | … |




| A | Yes | No |
|---|---|---|
| Yes | … | … |
| No | … | … |

| A | Yes | No |
|---|---|---|
| Yes | … | … |
| No | … | … |

**Structure Learning → Parameter Learning → Inference**

Imperial College London

## Project Aims and Deliverables

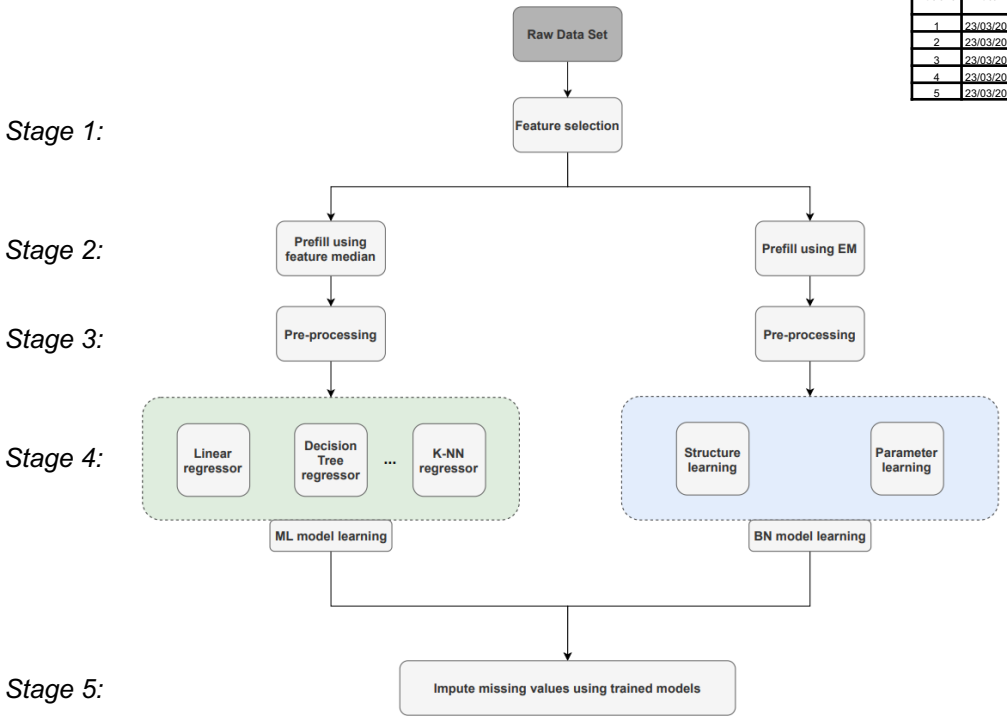
***Aim:*** *"Compare the performance of ML based and BN methods with simple median imputation."*

***Key Deliverables:***

- **Design:** Imputation framework which provides a *methodology*.
- **Implementation:** Develop the framework and create an *open source* library.
- **Experimentation:** Carry out an *empirical* study on a real-life laboratory data set.
- **Documentation:** Create a *reference* point for project findings and results.

Imperial College London

## Overview of Design & Implementation




| Patient | Date | Lab Code | | | | |
|---|---|---|---|---|---|---|
| | | EOS | MONO | BASO | .... | NRBCA |
| 1 | 23/03/20 | 0.40 | NaN | 0.10 | .... | 0.00 |
| 2 | 23/03/20 | 0.20 | NaN | 0.10 | .... | 0.00 |
| 3 | 23/03/20 | 0.40 | 0.50 | NaN | .... | 0.00 |
| 4 | 23/03/20 | 0.10 | NaN | 0.10 | .... | NaN |
| 5 | 23/03/20 | NaN | 0.40 | 0.10 | .... | 0.00 |

| EOS | MONO | BASO | .... | NRBCA |
|---|---|---|---|---|
| 0.40 | NaN | 0.10 | .... | 0.00 |
| 0.20 | NaN | 0.10 | .... | 0.00 |
| 0.40 | 0.50 | NaN | .... | 0.00 |
| 0.10 | NaN | 0.10 | .... | NaN |
| NaN | 0.40 | 0.10 | .... | 0.00 |

| EOS | MONO | BASO | .... | NRBCA |
|---|---|---|---|---|
| 0.40 | 0.47 | 0.10 | .... | 0.00 |
| 0.20 | 0.47 | 0.10 | .... | 0.00 |
| 0.40 | 0.50 | 0.10 | .... | 0.00 |
| 0.10 | 0.47 | 0.10 | .... | 0.00 |
| 0.28 | 0.40 | 0.10 | .... | 0.00 |

| EOS | MONO | BASO | .... | NRBCA |
|---|---|---|---|---|
| 0.40 | 0.47 | 0.10 | .... | 0.00 |
| 0.40 | 0.50 | 0.10 | .... | 0.00 |
| 0.28 | 0.40 | 0.10 | .... | 0.00 |

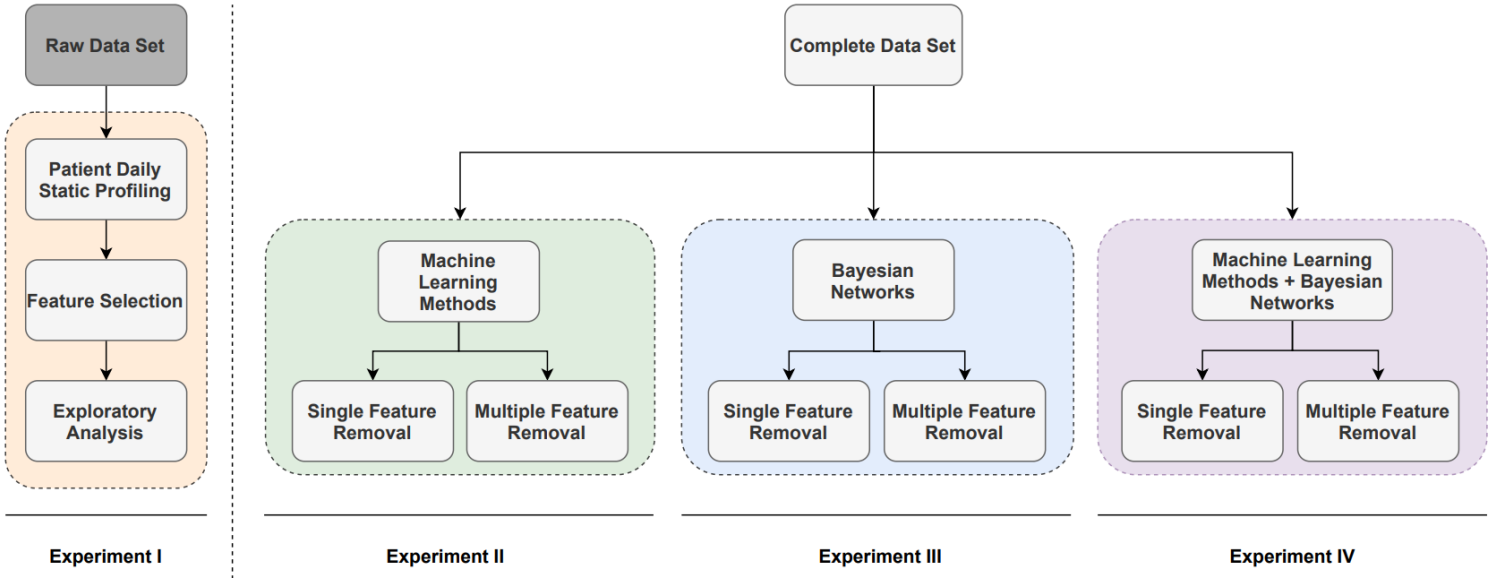
| EOS | MONO | BASO | .... | NRBCA |
|---|---|---|---|---|
| 0.40 | 0.47 | 0.10 | .... | 0.00 |
| 0.40 | 0.50 | 0.10 | .... | 0.00 |
| 0.28 | 0.40 | 0.10 | .... | 0.00 |

| EOS | MONO | BASO | .... | NRBCA |
|---|---|---|---|---|
| 0.40 | 0.38 | 0.10 | .... | 0.00 |
| 0.40 | 0.50 | 0.08 | .... | 0.00 |
| 0.21 | 0.40 | 0.10 | .... | 0.00 |

Full description of the analysis, design and implementation details in Chapter 4 and 5 of Final Report.

Imperial College London

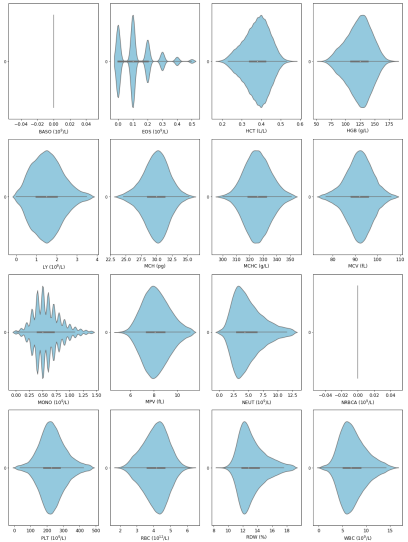
# Experiments

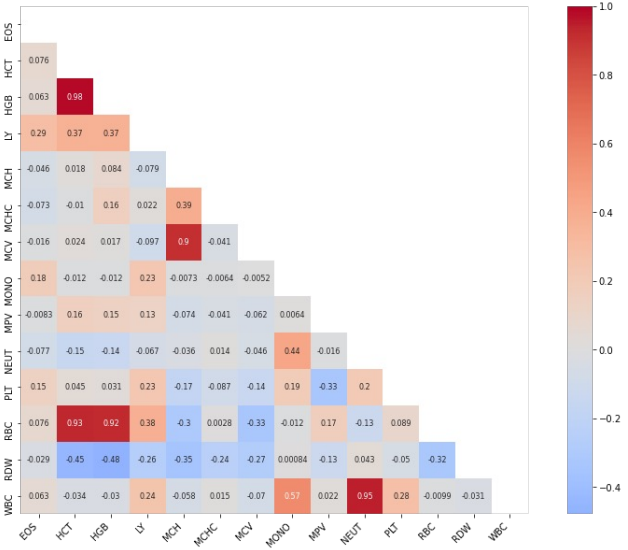


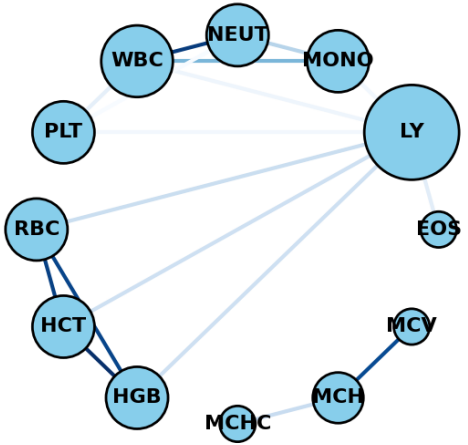

Full description of the experiment methodology presented in Chapter 6 of Final Report.

Imperial College London

# Experiment I – Overview of pathology data






Full description of results available in Chapter 7 – Section 2 of Final Report.

Imperial College London

# Results: Evaluation Metrics

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\Delta = 100 - \left(100 \times \frac{RMSE_{ML/BN}}{RMSE_{median}}\right)$$

$$NAE = \frac{|y - y_i|}{max(y_i) - min(y_i)}$$

Full description of evaluation metrics presented in Chapter 7 – Section 7.1.

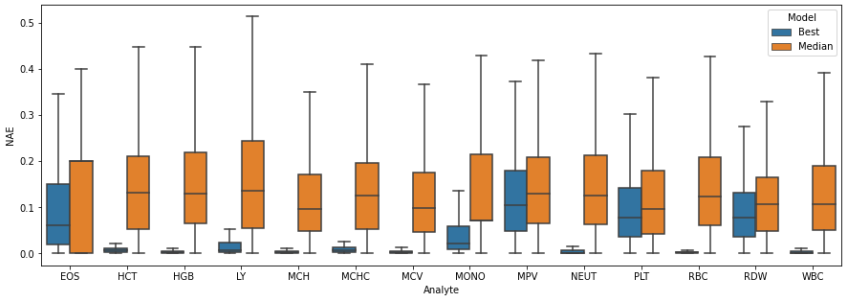
Imperial College London

# Experiment II – Imputation using ML based methods

## Single Feature Removal

For all analytes, error distribution for ML based (orange) is significantly **lower** than median imputation.

→ *ML methods impute with **higher** accuracy than median.*

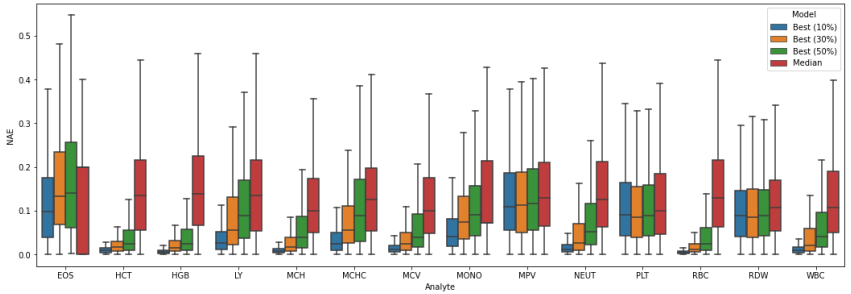



## Multiple Feature Removal

For all proportions of missing values (10%, 30% and 50%), central *tendency* and *dispersion* remains **below** median imputation.

→ *median **changes** the distribution of data.*

→ *ML methods to a much **lower** extent (relatively).*

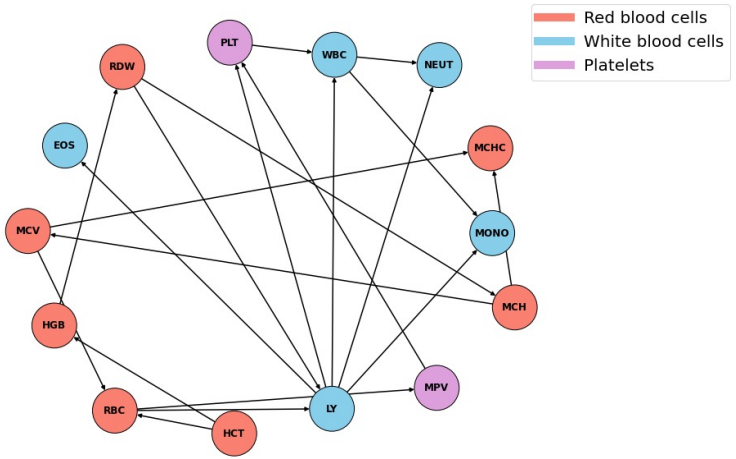



Remaining results presented in Chapter 7 – Section 7.2 of Final Report.

Imperial College London

# Experiment III – Imputation using BNs (1)

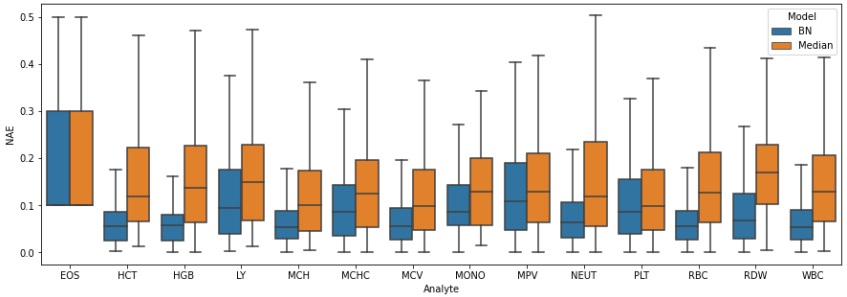
## Single Feature Removal




## Single Feature Removal

For all analytes, error distribution for BNs (blue) is **lower** than median imputation (orange).

→ *BNs perform **better** than median imputation.*

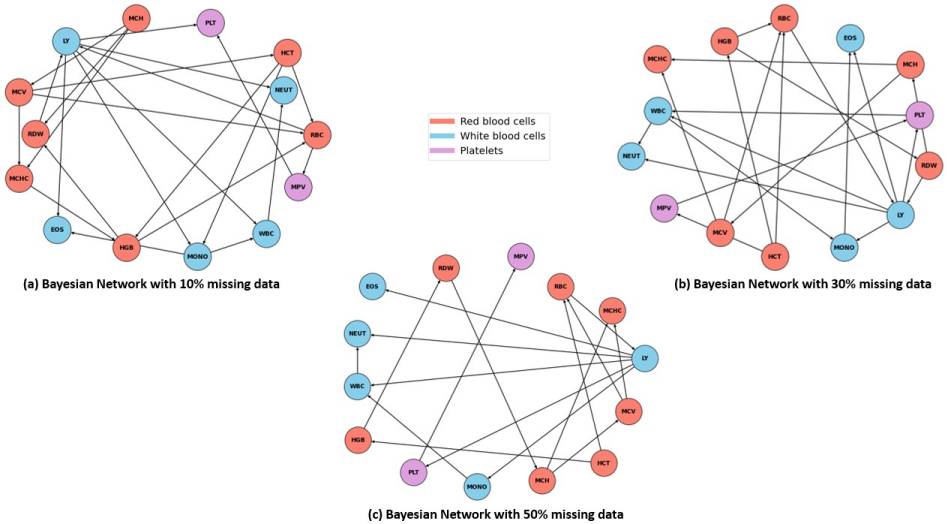



Remaining results presented in Chapter 7 – Section 7.3 of Final Report.

Imperial College London

# Experiment III – Imputation using BNs (2)

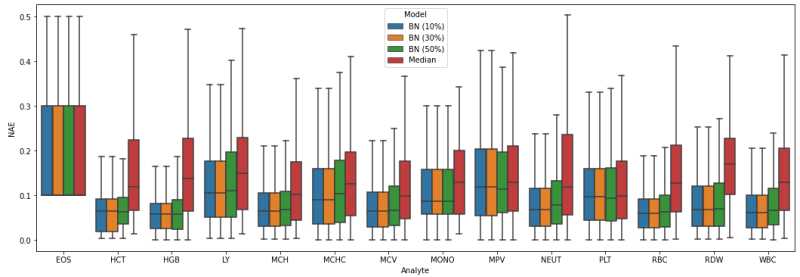
## Multiple Feature Removal




## Multiple Feature Removal

For all proportions of missing values (10%, 30% and 50%), central *tendency* and *dispersion* remains **below** median imputation.

→ *BN central tendencies are heterogenous so they are **more** consistent.*




Remaining results presented in Chapter 7 – Section 7.3 of Final Report.

Imperial College London

# Experiment IV – Comparing ML based and BN methods

## Single Feature Removal

RMSE scores are more **comparable**.

→ *discretisation of data **impacts** performance of ML methods.*

| Analyte | Unit | Machine Learning (ML) | | | Bayesian Network (BN) | |
|---|---|---|---|---|---|---|
| | | Type | HOTS | CVTS | HOTS | CVTS |
| EOS | $10^9$/L | **LR** | 0.108 | 0.107 | **0.103** | 0.102 |
| HCT | L/L | **LR** | **0.017** | 0.017 | 0.025 | 0.027 |
| HGB | g/L | **MLP** | **5.784** | 5.863 | 8.380 | 8.372 |
| LY | $10^9$/L | **LR** | 0.591 | 0.536 | **0.587** | 0.590 |
| MCH | pg | **MLP** | **0.624** | 0.624 | 0.919 | 0.920 |
| MCHC | g/L | **MLP** | **6.625** | 6.598 | 7.084 | 7.080 |
| MCV | fL | **MLP** | **2.061** | 2.056 | 2.749 | 2.733 |
| MONO | $10^9$/L | **LR** | **0.202** | 0.200 | 0.217 | 0.215 |
| MPV | fL | **MLP** | **0.993** | 0.998 | 1.036 | 1.033 |
| NEUT | $10^9$/L | **LR** | **0.872** | 0.844 | 1.137 | 1.135 |
| PLT | $10^9$/L | **MLP** | **65.496** | 66.922 | 67.472 | 67.469 |
| RBC | $10^{12}$/L | **MLP** | **0.209** | 0.215 | 0.325 | 0.323 |
| RDW | % | **MLP** | **1.242** | 1.314 | 1.300 | 1.298 |
| WBC | $10^9$/L | **LR** | **0.744** | 0.716 | 1.186 | 1.184 |
| **Average** | - | **-** | **6.112** | 6.638 | 6.609 | 6.605 |

## Multiple Feature Removal

For all proportions of missing values (10%, 30% and 50%), more **variation** in the best method.

→ *performance of ML methods **degrades** linearly.*

→ *BNs have a **lower** RMSE at higher missing values.*

| Analyte | Unit | Missing | Machine Learning (ML) | | | Bayesian Network (BN) | |
|---|---|---|---|---|---|---|---|
| | | (%) | Method | HOTS | CVTS | HOTS | CVTS |
| EOS | $10^9$/L | 10 | MLP | **0.112** | 0.114 | **0.112** | 0.113 |
| | | 30 | MLP | **0.110** | 0.108 | 0.116 | 0.114 |
| | | 50 | RF | **0.114** | 0.116 | 0.117 | 0.115 |
| HCT | L/L | 10 | MLP | **0.018** | 0.019 | 0.026 | 0.025 |
| | | 30 | MLP | **0.027** | 0.030 | 0.033 | 0.032 |
| | | 50 | MLP | **0.038** | 0.039 | 0.040 | 0.037 |
| HGB | g/L | 10 | MLP | **5.939** | 5.935 | 9.148 | 9.149 |
| | | 30 | MLP | **8.836** | 8.832 | 10.684 | 10.687 |
| | | 50 | MLP | 12.598 | 12.599 | **11.480** | 11.475 |
| LY | $10^9$/L | 10 | MLP | **0.552** | 0.550 | 0.616 | 0.614 |
| | | 30 | MLP | **0.622** | 0.620 | 0.643 | 0.639 |
| | | 50 | MLP | 0.690 | 0.698 | **0.676** | 0.675 |
| MCH | pg | 10 | MLP | **0.713** | 0.708 | 1.063 | 1.062 |
| | | 30 | MLP | 1.330 | 1.329 | **1.248** | 1.247 |
| | | 50 | MLP | 1.961 | 1.964 | **1.449** | 1.446 |
| MCHC | g/L | 10 | MLP | **7.132** | 7.130 | 7.684 | 7.682 |
| | | 30 | MLP | 8.966 | 8.962 | **8.544** | 8.542 |
| | | 50 | MLP | 10.793 | 10.791 | **9.045** | 9.043 |

*Best performing method is highlighted in **bold** showing it has a lower RMSE. Full table of results accessible in Chapter 7 – Section 7.4 of Final Report.

Imperial College London

# Conclusion, Achievements and Future Work

## Conclusions & Achievements

- Empirically shown simple median imputation performs **poorly**.
  → *high RMSE for all the analytes under all scenarios.*
  → *changes the distribution of underlying data.*
- Recommendation 1: ML based methods for CDSSs predictive modelling as they impute with **high** accuracy.
- Recommendation 2: BNs more suitable for clinicians as they are **interpretable** and **intuitive**.
- All project objectives were met.

## Future Work

- Integration into CDSS to carry out pilot studies.
- Extending experiments to consider other laboratory panels.
- Temporal profiling to use longitudinal data.
- Enhancing model performances especially BNs.