IVF data challenge

Marina Vabistsevits, Yi Liu, Peter Matthews

https://github.com/mvab/IVF-data-challenge

Dataset overview

HFEA - longest running register of fertility treatment data in the world

Main content:

- reasons for seeking treatment and obstetric history
- the type of treatment being used, the number of eggs collected, and the number of embryos transferred
- the number of babies born, their gestation and birth weight

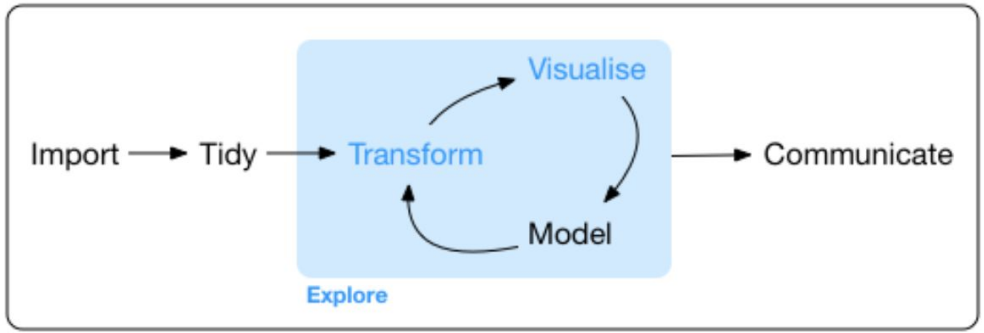
Current dataset:

- 1.37M observations
- data collected in 1999-2016
- ~100 variables for each patient

www.hfea.gov.uk

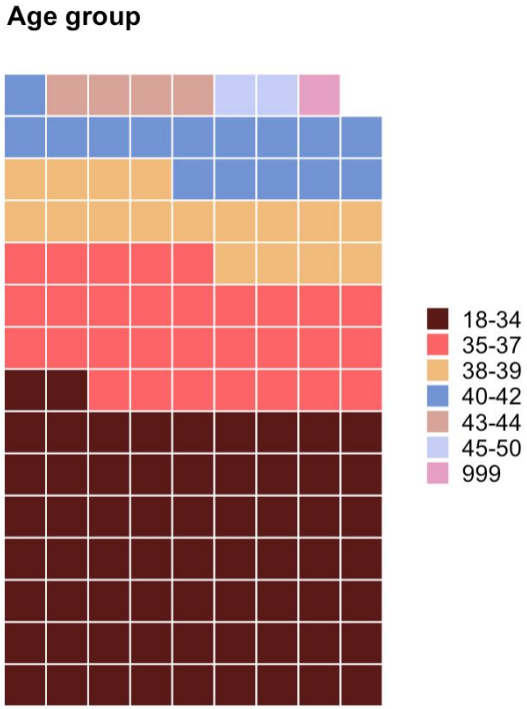
Plan / approach

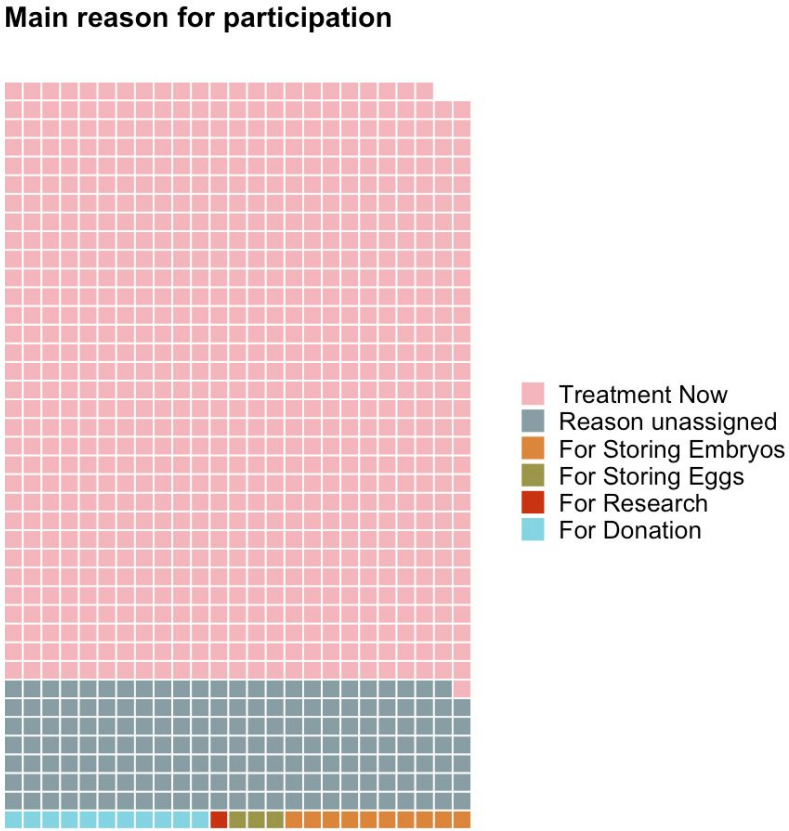
- ➔ Week 1: Exploratory Data Analysis
  - Variables
  - Missingness
  - Trends
  - + set questions for modelling
- ➔ Week 2: Modelling
  - Treatment success
  - Treatment choice



R for Data Science https://r4ds.had.co.nz/

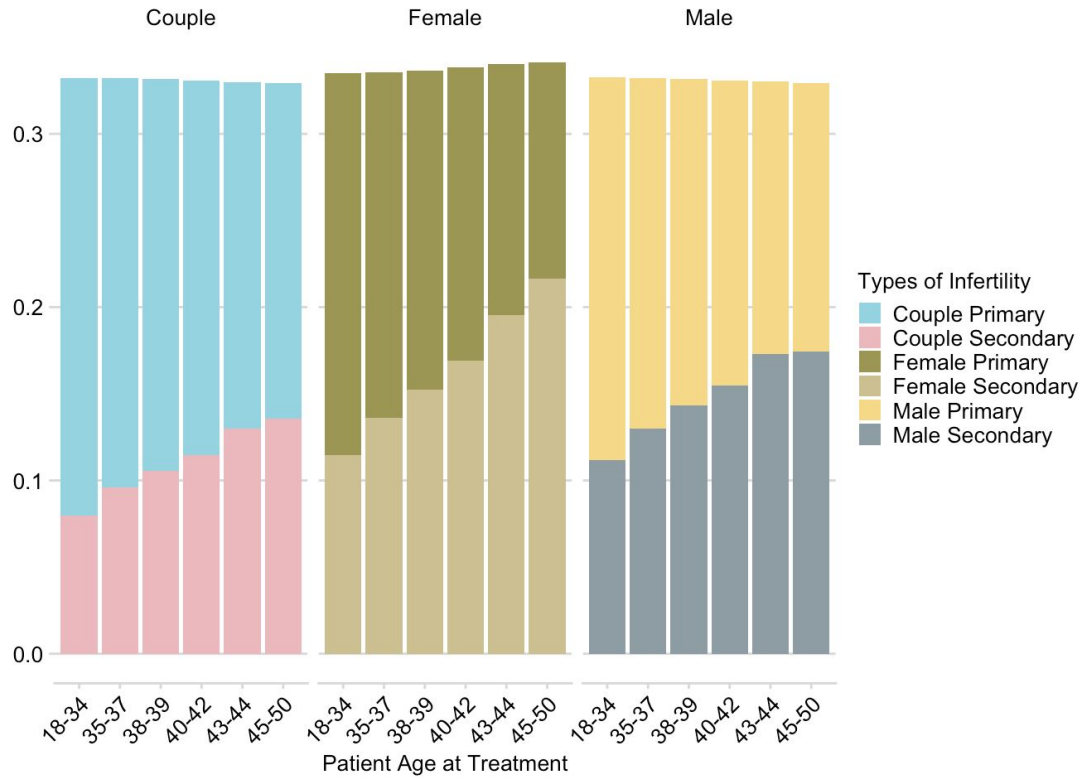
EDA1





https://github.com/mvab/IVF-data-challenge/blob/master/marina-working/ivf_eda.ipynb

EDA2 - types of infertility



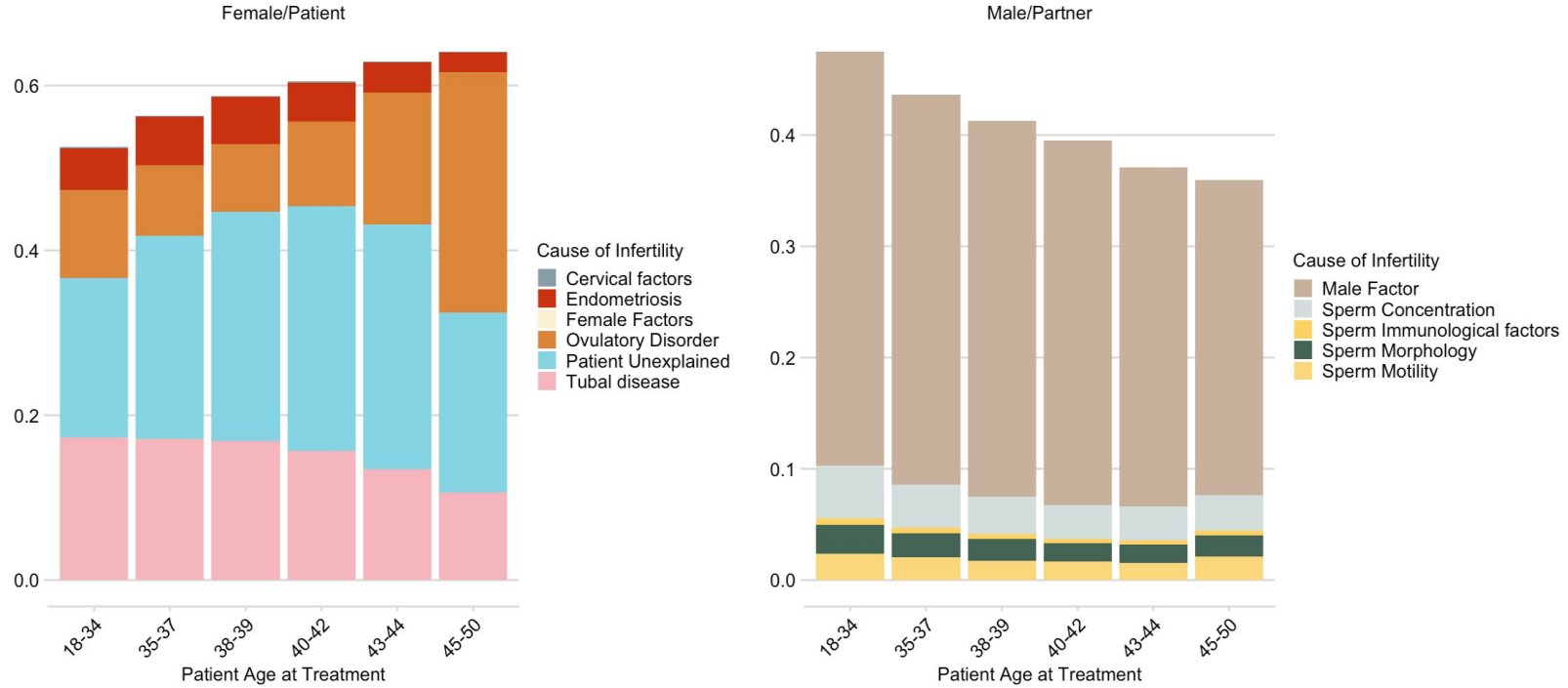
Infertility is usually only diagnosed when a couple have not managed to conceive after a year of trying.

There are 2 types of infertility:

- **primary infertility** – where someone who's never conceived a child in the past has difficulty conceiving
- **secondary infertility** – where someone has had 1 or more pregnancies in the past, but is having difficulty conceiving again

https://www.nhs.uk/conditions/infertility/

EDA3 - causes of infertility



EDA4 - treatment types background

Two main methods:

- **IVF (In Vitro Fertilisation)** - An egg being fertilised by sperm outside the body
- **DI (Donor Insemination)** - Using sperm from a sperm donor in order to get pregnant

In the dataset: IVF : 80% / DI: 20%

Many other subtypes within each category, most common:

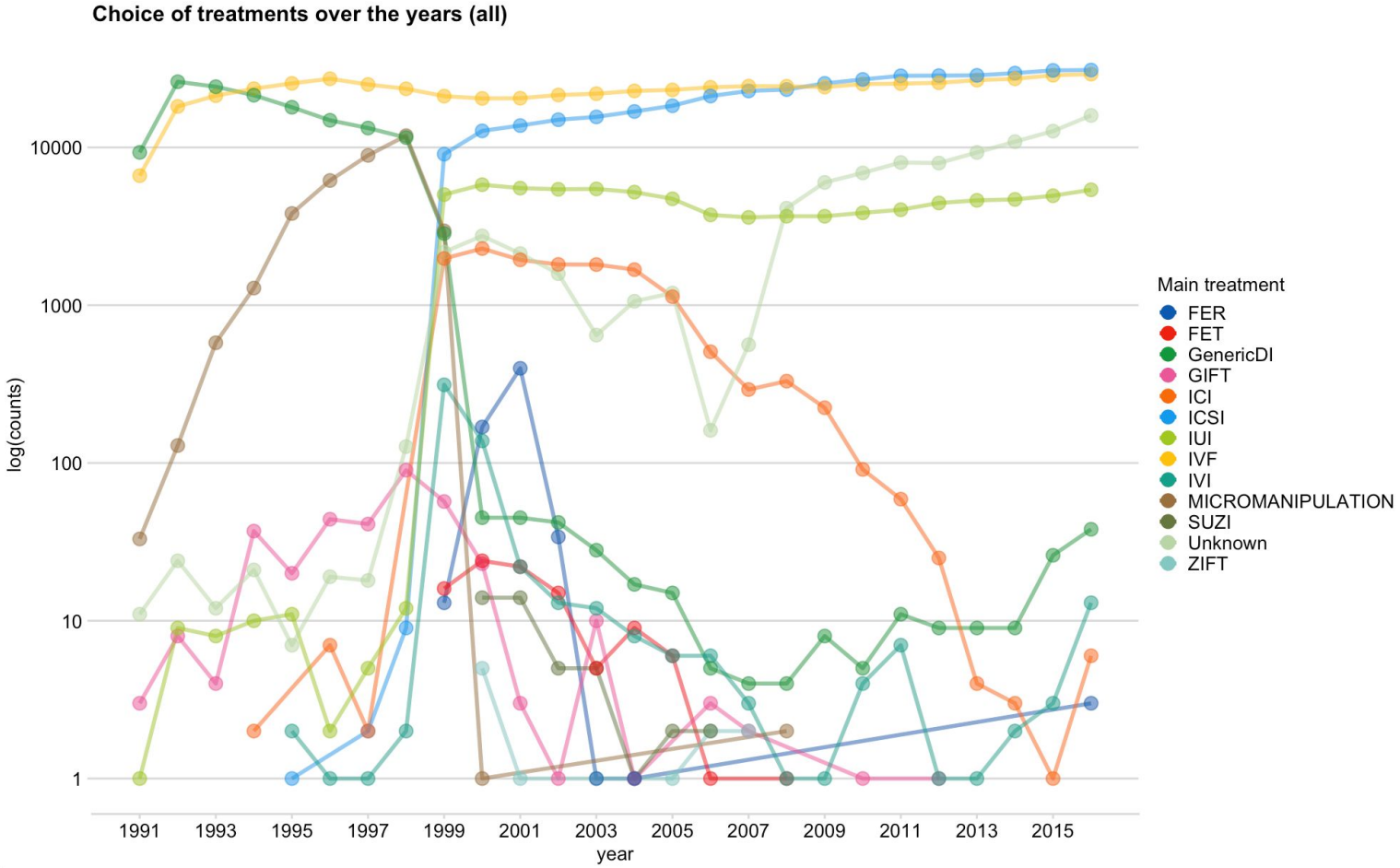
- **ICSI (Intracytoplasmic Sperm Injection)** - When a lone, high-quality sperm is injected straight into your egg during IVF, rather than allowing the sperm and egg to find one another in the dish

  - subtype of IVF (28%)

- **IUI (Intrauterine insemination)** involves directly inserting sperm into a woman's womb

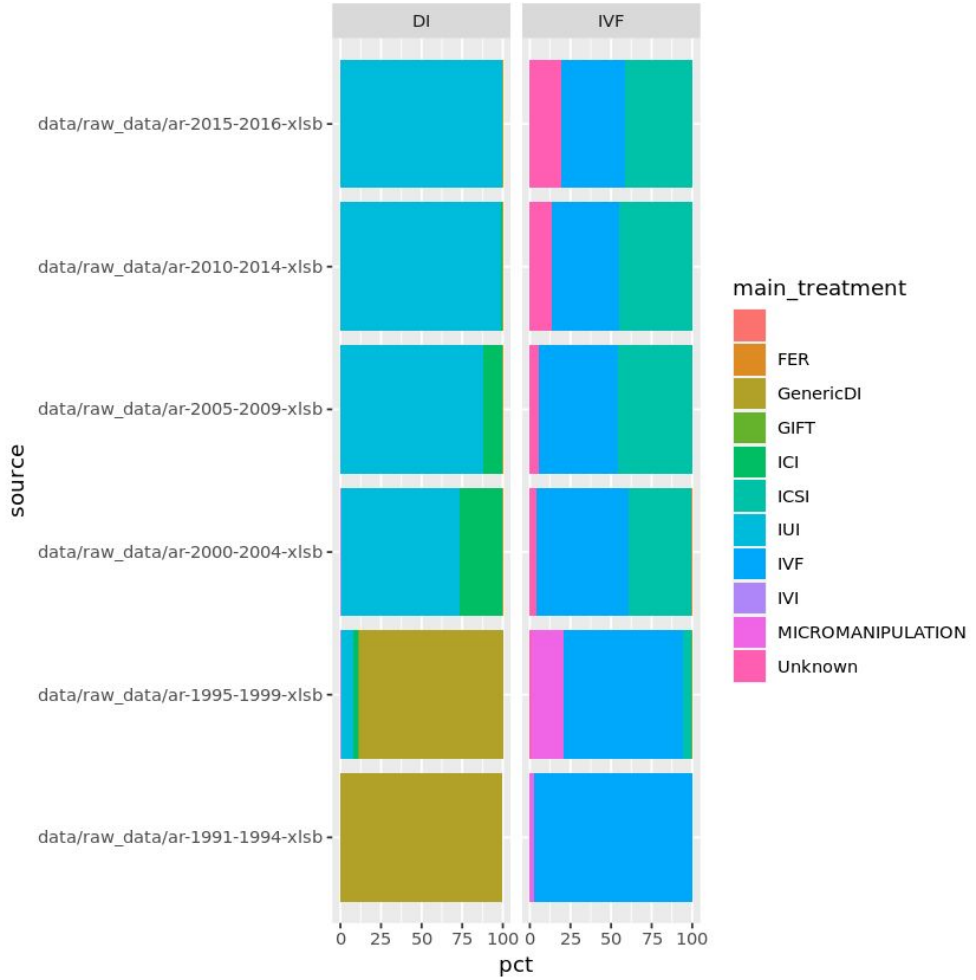
  - subtype of DI (6%)

EDA5


Choice of treatments over the years (all)
log(counts)
10000
1000
100
10
1
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015
year
Main treatment
FER
FET
GenericDI
GIFT
ICI
ICSI
IUI
IVF
IVI
MICROMANIPULATION
SUZI
Unknown
ZIFT

EDA 5 - structural break

- The pre-2000 cohorts contain distinct characteristics comparing to the post-2000 cohorts.
- The safest way is to treat the pre-2000 cohorts differently, and not conduct analysis on full sample.



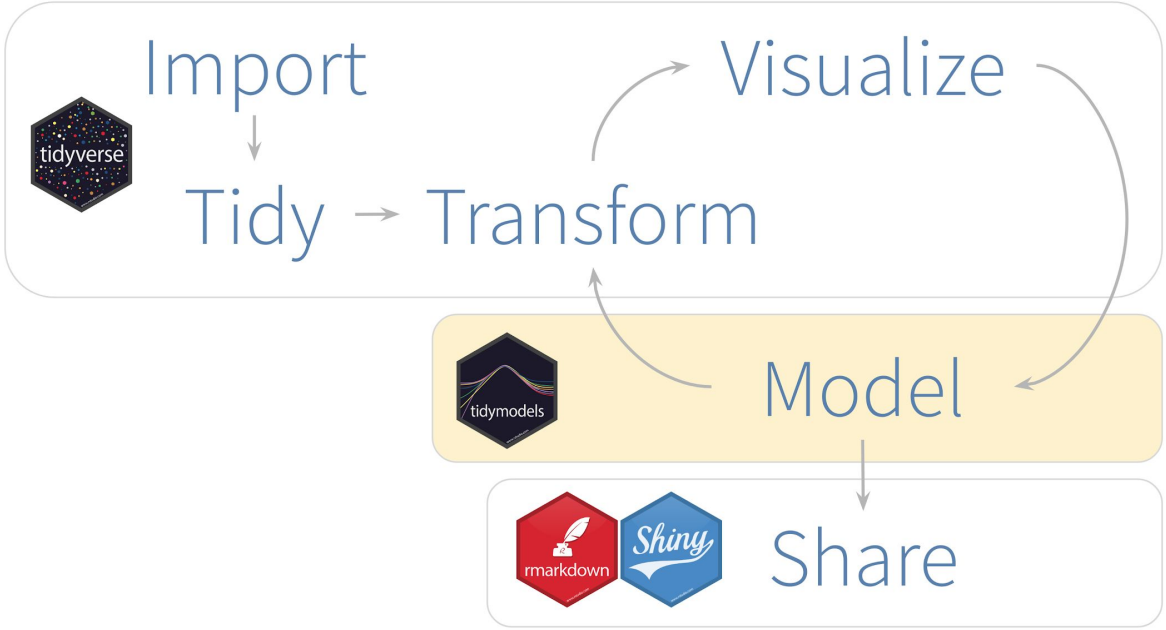
Modelling

2 types of predictions:

1. Type of treatment assigned to patient based on infertility/history
2. Treatment success : live birth occurrence

General rule:

- Standard ML approach: clean, split, preprocess, train, assess
- As little reliance on domain knowledge as possible
- Minimal, least invasive and rule based data cleaning and preprocessing
- Relying on training / testing results to assess over-fitting



tidymodels

- Successor to caret
- Sister package to tidyverse
- “One-stop-shop” for ML in R

Pre-Process → Train → Validate

rsample recipes parsnip yardstick

Live birth occurrence - introduction

IVFpredict (Nelson and Lawlor 2011)

- Logistic regression
- Predictors:
  - Maternal age
  - Duration infertility
  - Cause infertility
  - Num. unsuccessful IVF
  - Obstetric history
  - Hormonal preparation
  - Cycler number
  - Source of egg
  - Treatment type

OPEN ACCESS Freely available online PLoS MEDICINE

Predicting Live Birth, Preterm Delivery, and Low Birth Weight in Infants Born from In Vitro Fertilisation: A Prospective Study of 144,018 Treatment Cycles

Scott M. Nelson[1]*, Debbie A. Lawlor[2]*

1 Centre for Population and Health Sciences, University of Glasgow, Glasgow, Scotland, United Kingdom, **2** MRC Centre for Causal Analysis in Translational Epidemiology, School of Social and Community Medicine, University of Bristol, Bristol, England, United Kingdom

Abstract

Background: The extent to which baseline couple characteristics affect the probability of live birth and adverse perinatal outcomes after assisted conception is unknown.

Methods and Findings: We utilised the Human Fertilisation and Embryology Authority database to examine the predictors of live birth in all in vitro fertilisation (IVF) cycles undertaken in the UK between 2003 and 2007 ($n = 144,018$). We examined the potential clinical utility of a validated model that pre-dated the introduction of intracytoplasmic sperm injection (ICSI) as compared to a novel model. For those treatment cycles that resulted in a live singleton birth ($n = 24,226$), we determined the associates of potential risk factors with preterm birth, low birth weight, and macrosomia. The overall rate of at least one live birth was 23.4 per 100 cycles (95% confidence interval [CI] 23.2–23.7). In multivariable models the odds of at least one live birth decreased with increasing maternal age, increasing duration of infertility, a greater number of previously unsuccessful IVF treatments, use of own oocytes, necessity for a second or third treatment cycle, or if it was not unexplained infertility. The association of own versus donor oocyte with reduced odds of live birth strengthened with increasing age of the mother. A previous IVF live birth increased the odds of future success (OR 1.58, 95% CI 1.46–1.71) more than that of a previous spontaneous live birth (OR 1.19, 95% CI 0.99–1.24); p-value for difference in estimate <0.001. Use of ICSI increased the odds of live birth, and male causes of infertility were associated with reduced odds of live birth only in couples who had not received ICSI. Prediction of live birth was feasible with moderate discrimination and excellent calibration; calibration was markedly improved in the novel compared to the established model. Preterm birth and low birth weight were increased if oocyte donation was required and ICSI was not used. Risk of macrosomia increased with advancing maternal age and a history of previous live births. Infertility due to cervical problems was associated with increased odds of all three outcomes—preterm birth, low birth weight, and macrosomia.

Conclusions: Pending external validation, our results show that couple- and treatment-specific factors can be used to provide infertile couples with an accurate assessment of whether they have low or high risk of a successful outcome following IVF.

Please see later in the article for the Editors' Summary.

Live birth occurrence - cleaning & preprocessing

Cleaning

- Remove pre-2000 cohorts (see EDA)
- Remove variables that mechanistically predict birth outcome
  - “birth”, “foetus”, “early outcome”

Preprocessing

- Rebalance outcome classes (only on training set)
- Dummy encoding of categorical variables
- Remove predictors with near-zero-variance for models to stay parsimonious

https://github.com/mvab/IVF-data-challenge/blob/master/yi-working/birth_minimal.ipynb

- Raw: 1,376,454 rows; 96 cols;
- Cleaned: 933,358 rows; 62 cols
- Training set:
  - Overall: 314,525 rows; 87 cols (expanded dummies)
  - Outcome: 0: 152584; 1: 161,941
- Testing set:
  - Overall: 170,132 rows; 87 cols
  - Outcome: 0: 129,225; 1: 40907


```
── Data Summary ────────────────────────
                           Values
Name                       Piped data
Number of rows             314525
Number of columns          87
_____________________
Column type frequency:
  factor                   1
  numeric                  86
_____________________
Group variables            None

── Variable type: factor ──────────────────────────────────────────────────────
  skim_variable     n_missing complete_rate ordered n_unique
1 live_birth_occurr         0             1 FALSE          2
  top_counts
1 1: 161941, 0: 152584

── Variable type: numeric ─────────────────────────────────────────────────────
   skim_variable
 1 total_number_of_live_births___conceived_through_ivf
 2 type_of_infertility___female_primary
 3 type_of_infertility___female_secondary
 4 type_of_infertility___male_primary
 5 type_of_infertility___male_secondary
 6 type_of_infertility__couple_primary
 7 type_of_infertility__couple_secondary
 8 cause__of_infertility___tubal_disease
 9 cause_of_infertility___ovulatory_disorder
10 cause_of_infertility___male_factor
11 cause_of_infertility___patient_unexplained
12 cause_of_infertility___endometriosis
13 stimulation_used
14 elective_single_embryo_transfer
15 fresh_cycle
16 frozen_cycle
17 embryos_transfered
18 embryos_transfered_from_eggs_micro_injected
19 year_of_treatment
20 treatment_ivf_ivf
21 treatment_ivf_icsi
22 treatment_ivf_unknown
23 patient_age_at_treatment_X35.37
24 patient_age_at_treatment_X38.39
25 patient_age_at_treatment_X40.42
26 total_number_of_previous_cycles__both_ivf_and_di_X0
27 total_number_of_previous_cycles__both_ivf_and_di_X1
28 total_number_of_previous_cycles__both_ivf_and_di_X2
29 total_number_of_previous_cycles__both_ivf_and_di_X3
30 total_number_of_previous_cycles__both_ivf_and_di_other
31 total_number_of_previous_treatments__both_ivf_and_di_at_clinic_X1
32 total_number_of_previous_treatments__both_ivf_and_di_at_clinic_X2
33 total_number_of_previous_treatments__both_ivf_and_di_at_clinic_X3
34 total_number_of_previous_treatments__both_ivf_and_di_at_clinic_other
```

Live birth occurrence -- models

- Boosted tree (xgboost)
- Random forest (R ranger)
- Decision tree (R C5.0)
- Penalised logistic regression (R glmnet)

https://github.com/mvab/IVF-data-challenge/blob/master/yi-working/

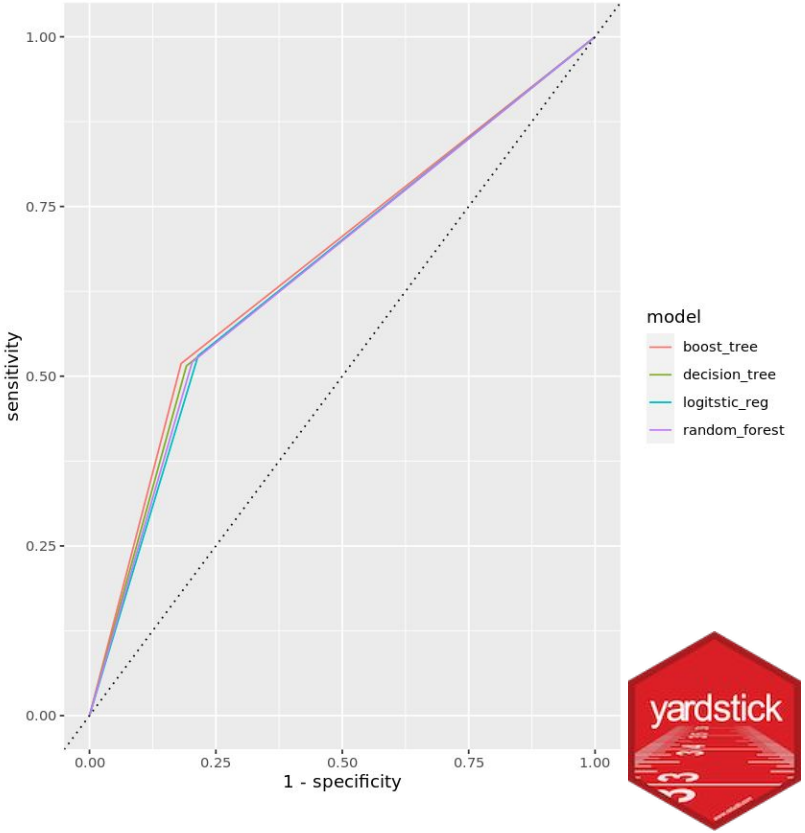
Live birth occurrence -- assessment

model	.metric	.estimator	.estimate
boost_tree	accuracy	binary	0.6797424
boost_tree	kap	binary	0.3535501
boost_tree	roc_auc	binary	0.6753093
decision_tree	accuracy	binary	0.6744483
decision_tree	kap	binary	0.3428825
decision_tree	roc_auc	binary	0.6700263
logitstic_reg	accuracy	binary	0.6623691
logitstic_reg	kap	binary	0.3193420
logitstic_reg	roc_auc	binary	0.6585507
random_forest	accuracy	binary	0.7790087
random_forest	kap	binary	0.5545862
random_forest	roc_auc	binary	0.7754060

Metrics on training set

model	.metric	.estimator	.estimate
boost_tree	accuracy	binary	0.5905651
boost_tree	kap	binary	0.2309591
boost_tree	roc_auc	binary	0.6687149
decision_tree	accuracy	binary	0.5855960
decision_tree	kap	binary	0.2217399
decision_tree	roc_auc	binary	0.6616971
logitstic_reg	accuracy	binary	0.5911783
logitstic_reg	kap	binary	0.2191906
logitstic_reg	roc_auc	binary	0.6576789
random_forest	accuracy	binary	0.5865415
random_forest	kap	binary	0.2179395
random_forest	roc_auc	binary	0.6588785

Metrics on testing set



Treatment type prediction

Predictions:

Binary classification: IVF vs DI

Multiclass prediction: IVF, ICSI, IUI, ICI

Models:

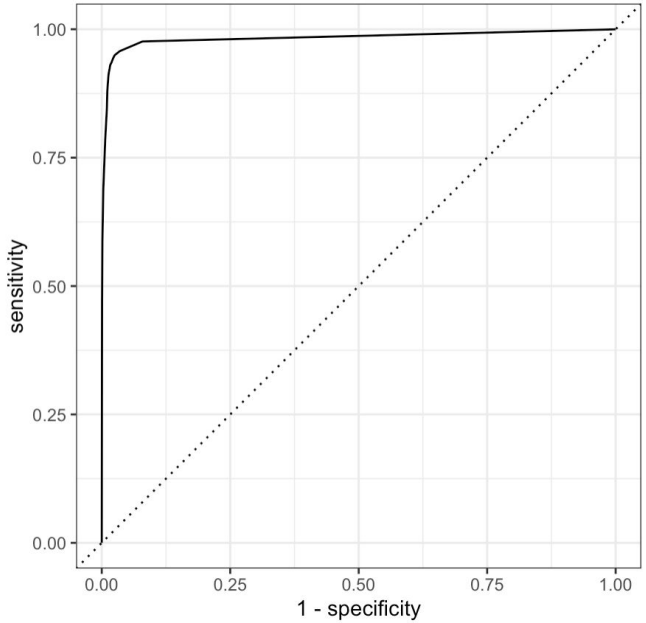
- *Random Forest*
- Logistic Regression
- xgboost

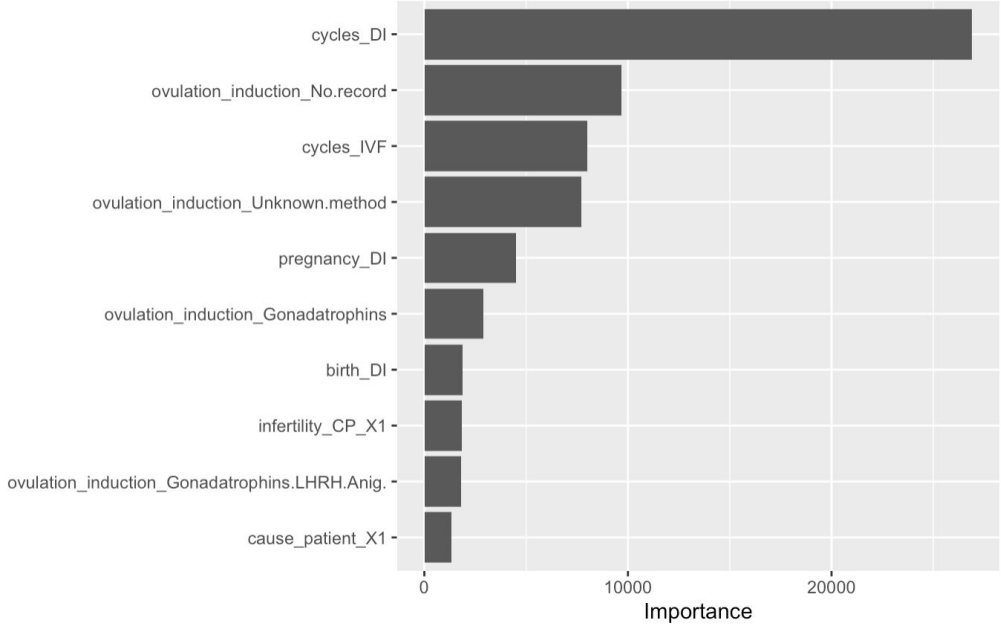
Included variables:

- Age group
- Types of infertility
  - Female/Male/Couple + Primary/Secondary
- Causes of infertility
  - Various female and male factors
- Patient history - Number of IVF/DI
  - cycles
  - pregnancies
  - live births
- Types of ovulation induction used

Treatment type (binary)

Including all variables: Accuracy: 0.98 / kappa: 0.88 / AUC ROC: 0.94

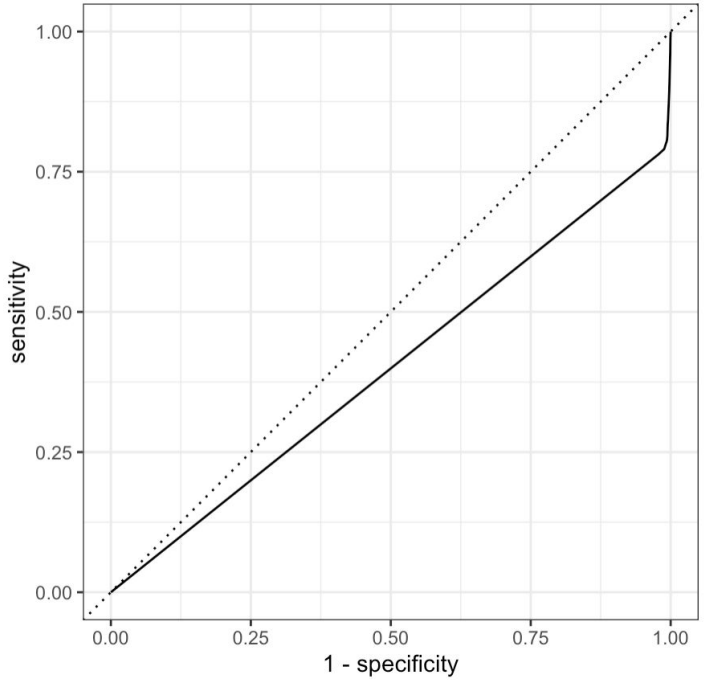


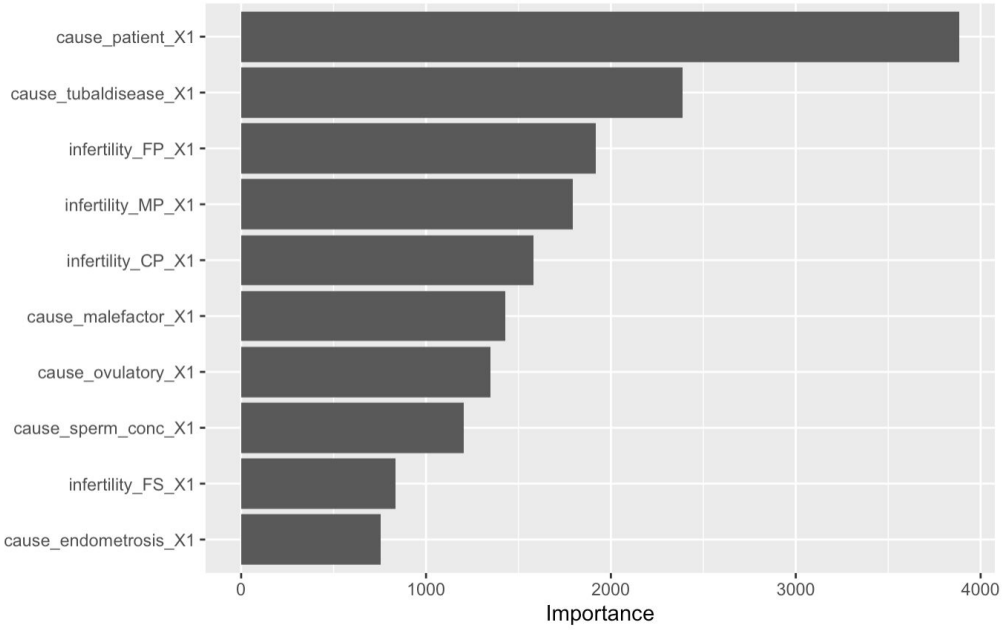


Treatment type (binary)

Including only age and infertility variables:

Accuracy: 0.91 / kappa: 0.27 / AUC ROC: 0.40

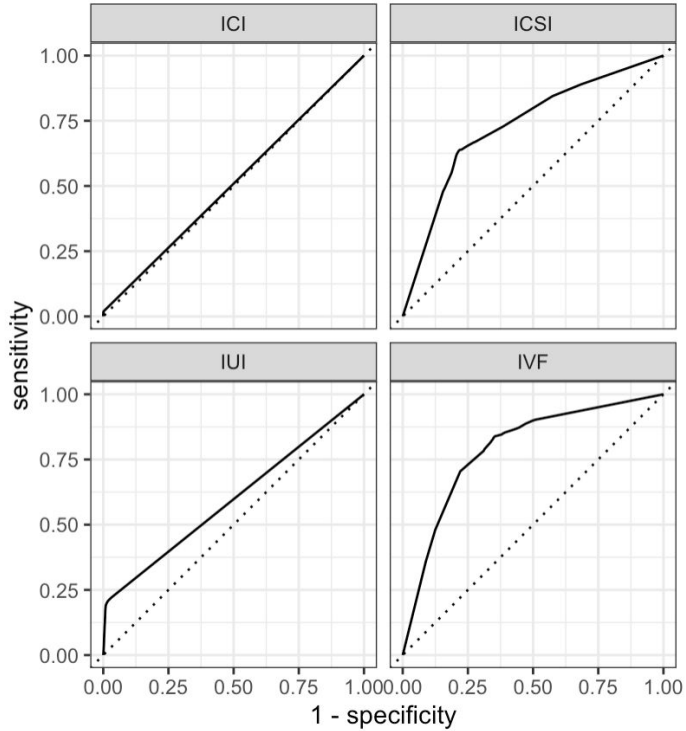


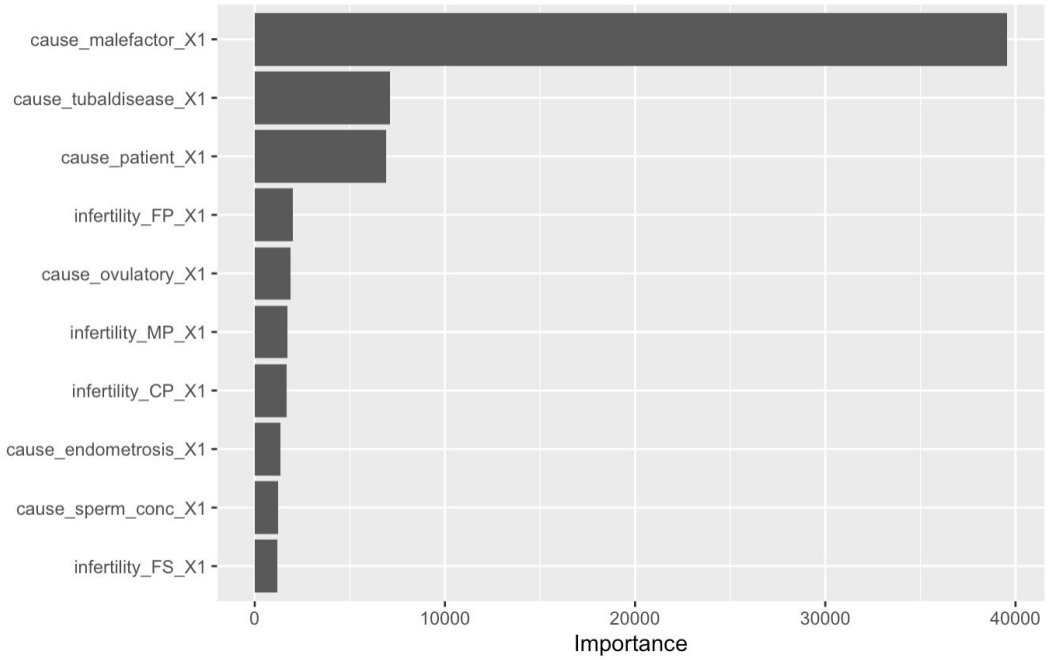


Treatment type (multiclass)

Including only age and infertility variables:

Accuracy: 0.67 / AUC ROC: 0.60





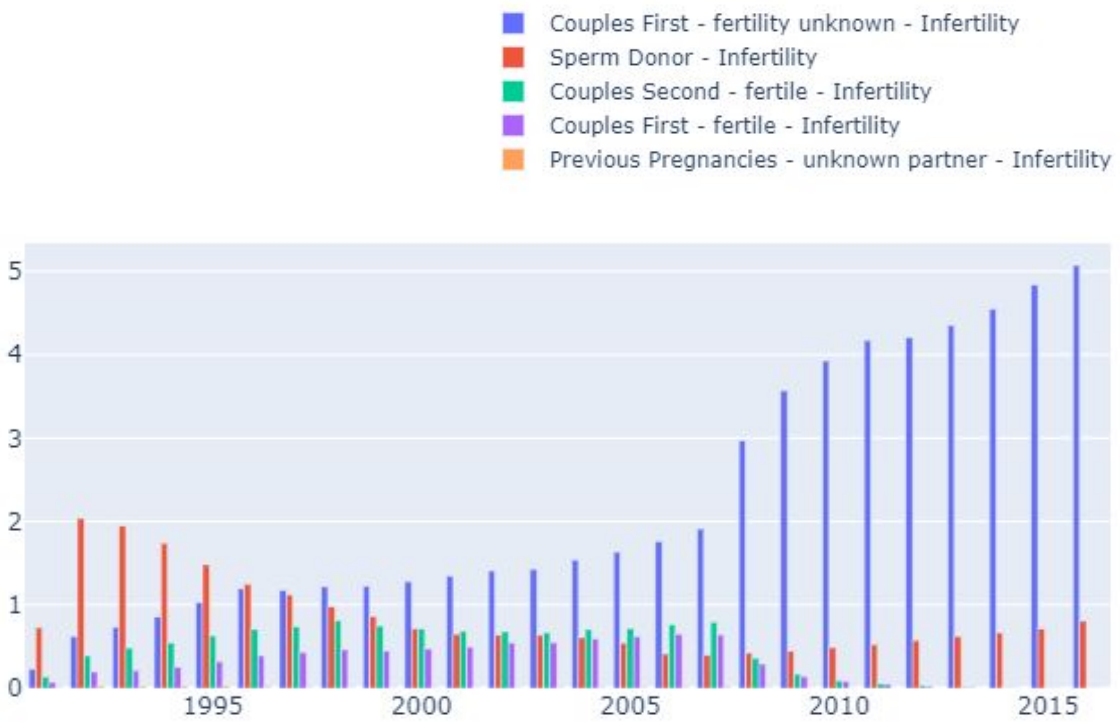
Data changes over the years

Missing or data collection changes:

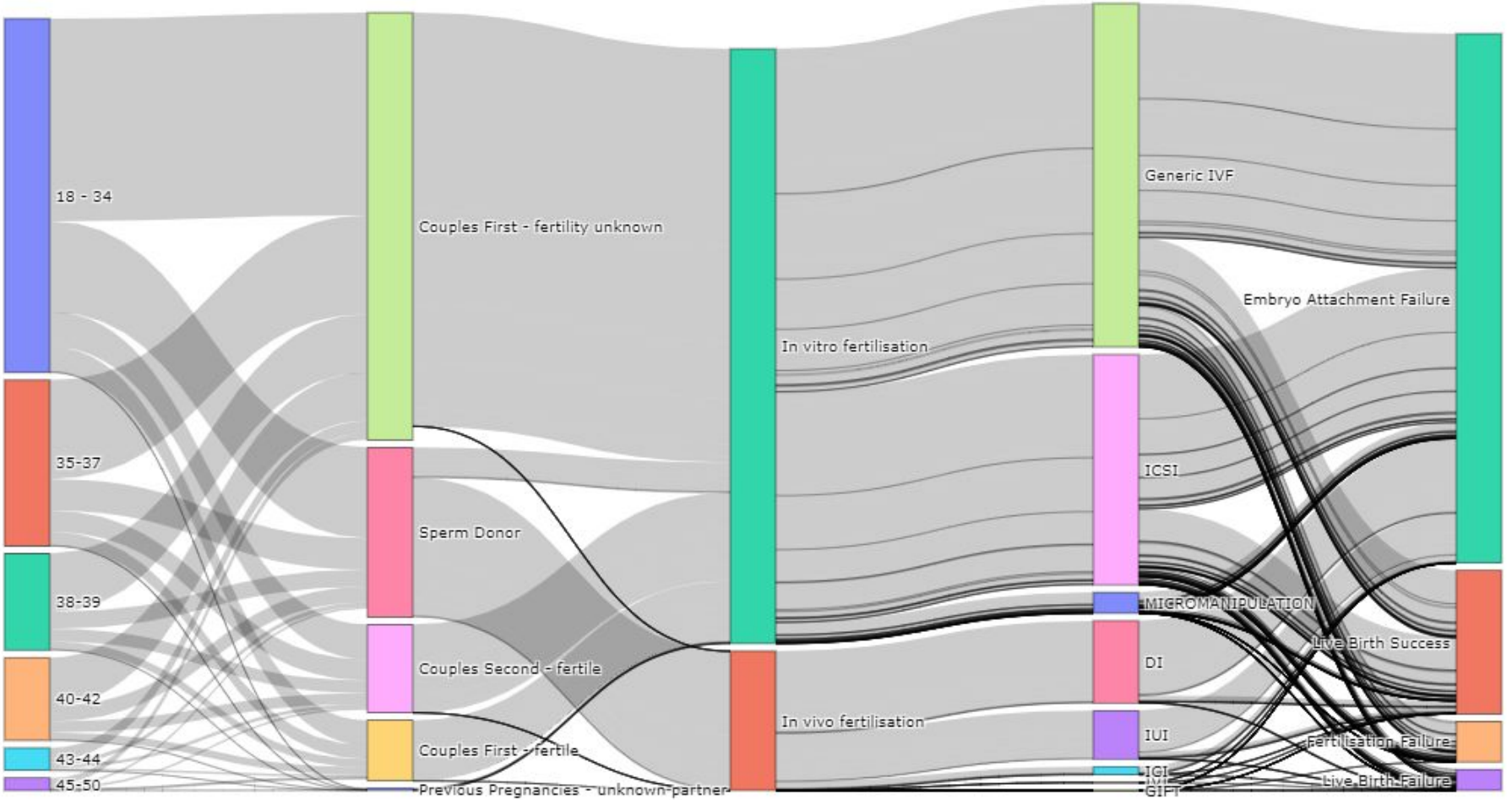
- Type of Infertility - data missing/trailing off after 2009.
- 20% of treatment classifications missing
- Post 1999 DI is sub classified into IUI, ICI, IVI and GIFT.

Methodology Changes:

- Post 2008 IVF methods such as ICS prefered over DI methods.
- IVF egg fertilization failures dramatically reduced by using ICSI method.



Data Visualisation of the HFEA data



Limitations

- Add limitations here
- Domain knowledge driven data cleaning & feature engineering
- Cross validation
- Hyper-parameter tuning
- Ensemble learning / super learner


Human Fertilisation &
Embryology Authority


R


tidyverse


tidymodels


pandas


rmarkdown


jupyter


plotly


dmlc
XGBoost