

Warfit Supplementary Material

Gianluca Truda

TRDGIA001

University of Cape Town

trdgia001@myuct.ac.za

Table 1: Comparison of raw IWPC and raw PathCare datasets in terms of key parameters.

Parameter	IWPC	PathCare
Number of records	6256	8983
Male fraction (%)	57	46
Mean age (years)	70-79	69.9
Mean INR (st.dev)	2.37 (0.46)	2.47 (0.94)
Mean weekly dose (mg)	31.6 (17.1)	28.8 (16.7)

ESTIMATION OF HUMAN PERFORMANCE

Experiment 1 compared the performance of clinicians to models on a subset of the PathCare dataset that contained patients with multiple visits (PATH-MV). Because no clinical experiment could be run whereby models dose real patients, we instead compared predictions to final therapeutic doses. To evaluate clinician performance, the first weekly dose prescribed to a patient was considered equivalent to a prediction of what the patient’s therapeutic dose should be. This was then compared to the final dose of the patient. If that final dose put their INR in the therapeutic range and was stable – meaning that the final taken dose and final prescribed dose differed by less than the standard deviation of weekly doses in the PATH-whole dataset – then the dose was considered the *true* dose. This was not necessarily the optimum dose for that patient, but was close enough to be used as a fair estimate. It should be noted that this biases the evaluation in favour of the human clinicians, not the models. The first dose prescribed was processed as the predictions (X values) and the final *true* dose was processed as the measured variable (Y values). As with any of the non-human models, these X and Y values were statistically analysed to find PW20, R^2 , and MAE scores.

Using the schema in table 2 as a reference: assume the final dose taken (D_{Tn}) puts the final INR value (INR_n) in the therapeutic range ($T \pm 0.5$). Assume also that the final dose taken (D_{Tn}) matched the second last dose prescribed ($D_{P_{n-1}}$). So, we can state:

$$(T - 0.5) \leq INR_n \leq (T + 0.5) \quad (1)$$

$$D_{P_{n-1}} = D_{Tn} \quad (2)$$

Now, a human (H) predicted the therapeutic dose (D'_T) that resulted in the therapeutic INR value (INR'), using the last dose taken ($D_{T_{n-1}}$), the last INR value (INR_{n-1}), the target INR (T) and the patient parameters. We express this as:

$$H(D_{T_{n-1}}, INR_{n-1}, T, Parameters) = D'_T = D_{Tn} \quad (3)$$

Because we cannot test the predictive accuracy of an algorithm (A) directly on a subject, the best we can do is compare the algorithm’s predicted dose (D_A) to the demonstrably-accurate human dose (D'_T) using the same parameters.

In cases where the patient’s final prescribed dose was either non-therapeutic or unstable, the *true* dose could not be inferred. This accounted for ~40% of the dataset. It would be unfair to compare only the best results of clinicians with all the results of models. For that reason, those doses were estimated. A fair way of doing this was to randomly sample from a normal distribution of weekly doses generated using the mean and standard deviation of the weekly doses in the PathCare-whole dataset. The magnitude of the difference between the mean dose and the sampled dose was then added to (or subtracted from) the final taken dose of the patient, depending on whether the INR was below or above the therapeutic range.

```
def estimate_dose(last_dose):
    """Estimates true dose based on last dose"""

    mean_dose = 28.82 # miu
    dev_dose = 16.71 # sigma

    multi = 1 if score > 0 else -1

    gaussian = random.gauss(mean_dose, dev_dose)
    dose = max(last_dose + multi * abs(mean_dose - gaussian), 0)

    return dose
```

This new value was considered a reasonable estimate of a *true* dose. For an individual case, the estimated value was likely different to the optimum dose for that patient. However, when repeating this process on all the missing values, the differences averaged out within the normal distribution of doses. Whilst it must be emphasised that this produced only an *estimate*, it is based on reasonable methodology. These estimates may not reflect the real-world dosing aptitude of the clinicians – for instance due to erroneous database records. Auditing the performance of clinicians was not the purpose of this study. The relation between the estimated performance of clinicians and models was the only concern, and since both the clinicians and models are evaluated on the same (possibly incomplete or erroneous dataset) the contest can be deemed fair.

Table 2: Assume the following data in CSV for a single patient with pid = x

pid	age, sex, drugs, conditions, etc.	doseTakenWeek	inr	targetINR	dosePrescribedWeek
x	Params	D_{T1}	INR_1	T	D_{P1}
x	Params	D_{T2}	INR_2	T	D_{P2}
...
x	Params	$D_{T_{n-1}}$	INR_{n-1}	T	$D_{P_{n-1}}$
x	Params	D_{T_n}	INR_n	T	D_{P_n}

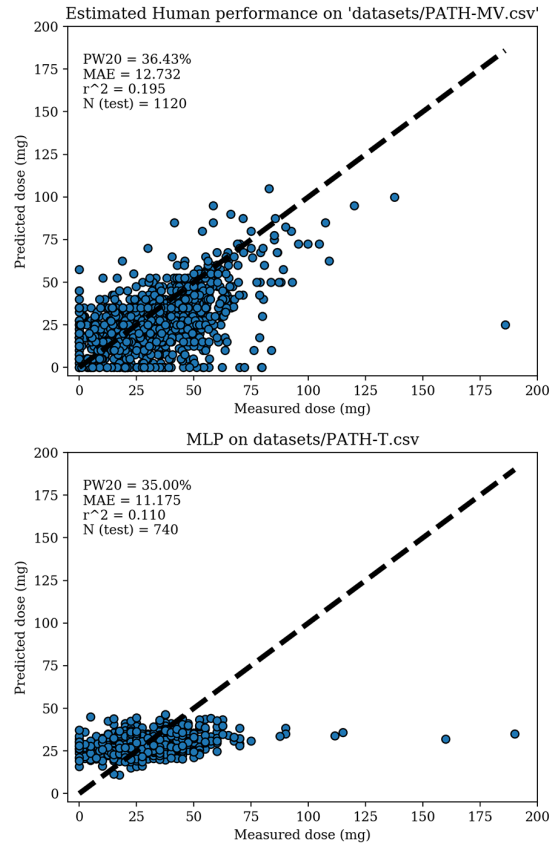


Figure 1: Comparison of estimated human dosing patterns with those of a multi-layer perceptron with tuned hyperparameters. The human dosing was estimated using PATH-MV. The MLP was trained on 80% of PATH-T with the 'path2' parameter set and validated on the remaining 20%. Whilst learning algorithms produce models that cluster predictions around the 20-40mg range, the human experts have a much more spread pattern, mapping closer to the $y = x$ line visually. Whilst the humans frequently prescribed doses lower than 20mg, this was uncommon for learned models. One explanation for this may be that human experts are typically more cautious when prescribing doses. The humans also appeared to predict doses about 75mg/week very successfully compared to the learned models. This could be, in part, due to additional information (provided in textual description) that was redacted from the training data, making it unavailable to the learning algorithms.

Terminal Dump from Full Experimental Run

```
gianlucatruda at gianluca in ~  
$ cd warfit
```

```
gianlucatruda at gianluca in ~/warfit on master [$]  
$ py main.py
```

```
EXP 1: Human vs. Algorithmic Dosing
```

```
PW20:      36.79 %  
MAE:      12.41  
r2-value:  0.209  
std. error: 0.020  
p-value: 1.221E-101  
N (test):  1120
```

```
1135 patients from 2792 rows captured, of which 1831 records are not therapeutic  
1120 patients dosed  
368 well dosed  
752 misdosed  
34 unsettled  
15 with too few sessions  
75 dosed 0mg
```

```
100.0%  
1135 patients from 2792 rows captured, of which 1831 records are not therapeutic  
1120 patients dosed  
368 well dosed  
752 misdosed  
34 unsettled  
15 with too few sessions  
75 dosed 0mg
```

```
* humans
```

```
PW20: 37.88% (35.89% - 39.46%)  
MAE: 12.672 (12.311 - 13.099)  
R^2: 0.187 (0.143 - 0.226)  
Time: 0ms (0ms - 0ms)
```

```
Loaded 'datasets/PATH-T.csv': 3696/3696 entries used, 9 features  
Loaded 'datasets/Withheld/PATH-V.csv': 925/925 entries used, 9 features  
Validating 13 models: validate-path-path2-mean
```

```
100.0%
```

```
* dLR
```

```
PW20: 34.59% (34.59% - 34.59%)  
MAE: 11.248 (11.248 - 11.248)  
R^2: 0.127 (0.127 - 0.127)  
Time: 1ms (1ms - 1ms)
```

```
* GG
```

```
PW20: 20.50% (18.27% - 22.49%)  
MAE: 17.894 (17.561 - 18.277)  
R^2: 0.000 (0.000 - 0.000)  
Time: 0ms (0ms - 0ms)
```

```
* RT
```

```
PW20: 35.14% (35.14% - 35.14%)  
MAE: 11.335 (11.335 - 11.335)  
R^2: 0.122 (0.122 - 0.122)  
Time: 2ms (1ms - 2ms)
```

```
* MLP
```

```
PW20: 34.75% (34.05% - 35.35%)  
MAE: 11.267 (11.242 - 11.292)  
R^2: 0.124 (0.116 - 0.129)  
Time: 1336ms (1015ms - 1714ms)
```

```
* SVR
```

```
PW20: 34.22% (33.41% - 34.92%)  
MAE: 11.340 (11.321 - 11.364)  
R^2: 0.094 (0.091 - 0.097)  
Time: 4ms (4ms - 5ms)
```

```
* LARS
```

```
PW20: 34.92% (34.92% - 34.92%)  
MAE: 11.235 (11.235 - 11.235)  
R^2: 0.127 (0.127 - 0.127)  
Time: 15ms (14ms - 16ms)
```

```
* PL1
```

PW20: 35.03% (35.03% - 35.03%)
MAE: 11.251 (11.251 - 11.251)
R^2: 0.127 (0.127 - 0.127)
Time: 22ms (22ms - 27ms)

* PL2
PW20: 35.37% (34.59% - 36.00%)
MAE: 11.235 (11.218 - 11.254)
R^2: 0.128 (0.126 - 0.130)
Time: 379ms (364ms - 391ms)

* PL3
PW20: 34.55% (34.05% - 34.92%)
MAE: 11.354 (11.332 - 11.386)
R^2: 0.097 (0.094 - 0.100)
Time: 269ms (252ms - 288ms)

* PL4
PW20: 35.29% (34.70% - 36.11%)
MAE: 11.235 (11.215 - 11.251)
R^2: 0.129 (0.127 - 0.131)
Time: 378ms (371ms - 387ms)

* PL5
PW20: 34.88% (34.49% - 35.24%)
MAE: 11.319 (11.279 - 11.356)
R^2: 0.102 (0.099 - 0.106)
Time: 203ms (200ms - 207ms)

* PL6
PW20: 34.81% (34.81% - 34.81%)
MAE: 11.231 (11.231 - 11.231)
R^2: 0.128 (0.128 - 0.128)
Time: 431ms (425ms - 442ms)

* PL7
PW20: 34.49% (34.49% - 34.49%)
MAE: 11.237 (11.237 - 11.237)
R^2: 0.125 (0.125 - 0.125)
Time: 37ms (36ms - 39ms)

13 rows written to 'results/exp1/path-by-alg_2018-09-12_12-38_val.csv'
Loaded 'datasets/RAW/PathCare/PATH-whole.csv': 8985/8985 entries used, 9 features
Validating 13 models: validate-path-path2-mean-k-fold
CV technique: k-fold
Comparing 13 models with 10-fold CV on validate-path-path2-mean-k-fold ...
N (per fold): 898

* dLR
PW20: 32.74% (30.51% - 36.37%)
MAE: 11.640 (10.820 - 12.275)
R^2: 0.123 (0.083 - 0.167)
Time: 2ms (2ms - 2ms)

* GG
PW20: 18.83% (15.37% - 20.71%)
MAE: 18.644 (17.824 - 19.755)
R^2: 0.000 (0.000 - 0.000)
Time: 0ms (0ms - 0ms)

* RT
PW20: 32.19% (29.29% - 34.26%)
MAE: 11.685 (11.162 - 12.397)
R^2: 0.120 (0.073 - 0.142)
Time: 3ms (3ms - 3ms)

* MLP
PW20: 32.50% (29.14% - 35.41%)
MAE: 11.650 (11.263 - 12.007)
R^2: 0.122 (0.080 - 0.192)
Time: 2270ms (1734ms - 2697ms)

* SVR
PW20: 32.27% (29.14% - 34.48%)
MAE: 11.593 (10.823 - 11.966)
R^2: 0.114 (0.098 - 0.145)
Time: 10ms (9ms - 12ms)

* LARS
PW20: 32.64% (29.48% - 34.59%)

MAE: 11.648 (11.095 - 12.571)
R²: 0.119 (0.084 - 0.149)
Time: 19ms (18ms - 20ms)

* PL1
PW20: 32.79% (30.96% - 34.97%)
MAE: 11.619 (11.149 - 12.025)
R²: 0.125 (0.101 - 0.156)
Time: 47ms (45ms - 50ms)

* PL2
PW20: 32.65% (31.29% - 33.59%)
MAE: 11.608 (11.096 - 12.340)
R²: 0.129 (0.096 - 0.156)
Time: 793ms (778ms - 814ms)

* PL3
PW20: 32.57% (30.73% - 34.26%)
MAE: 11.547 (11.050 - 12.059)
R²: 0.121 (0.091 - 0.152)
Time: 456ms (435ms - 466ms)

* PL4
PW20: 32.68% (29.96% - 36.08%)
MAE: 11.634 (11.229 - 12.042)
R²: 0.126 (0.090 - 0.155)
Time: 790ms (775ms - 818ms)

* PL5
PW20: 32.85% (30.92% - 35.15%)
MAE: 11.548 (11.124 - 12.357)
R²: 0.124 (0.082 - 0.158)
Time: 348ms (332ms - 367ms)

* PL6
PW20: 32.65% (30.40% - 33.85%)
MAE: 11.642 (11.115 - 12.090)
R²: 0.122 (0.090 - 0.168)
Time: 1055ms (1032ms - 1088ms)

* PL7
PW20: 32.50% (31.51% - 33.70%)
MAE: 11.717 (11.212 - 12.130)
R²: 0.111 (0.091 - 0.127)
Time: 55ms (49ms - 62ms)

13 rows written to 'results/exp1/path-by-alg_2018-09-12_12-38_xval.csv'

EXP 2A: Comparison of Models Across Parameter Sets.
Comparing parameter sets on 'datasets/Withheld/PATH-V.csv' ...

Parameter set: common, n_train = 3696, n_val = 925
100.0%
Parameter set: path1, n_train = 3696, n_val = 925
100.0%
Parameter set: path2, n_train = 3696, n_val = 925
100.0%

* common
PW20: 33.88% (32.76% - 34.81%)
MAE: 11.368 (11.249 - 11.683)
R²: 0.110 (0.083 - 0.126)
Time: 286ms (1ms - 2111ms)

* path1
PW20: 33.79% (30.05% - 36.32%)
MAE: 11.441 (11.194 - 12.244)
R²: 0.097 (0.000 - 0.127)
Time: 367ms (1ms - 2711ms)

* path2
PW20: 34.16% (32.22% - 36.22%)
MAE: 11.376 (11.198 - 11.864)
R²: 0.104 (0.030 - 0.130)
Time: 327ms (1ms - 2224ms)

3 rows written to 'results/exp2a/path-by-params_2018-09-12_12-38.csv'
Comparing parameter sets on 'datasets/Withheld/IWPC-V.csv' ...

Parameter set: common, n_train = 3623, n_val = 906

```

100.0%
Parameter set: clinical, n_train = 3623, n_val = 906
100.0%
Parameter set: liu, n_train = 3623, n_val = 906
100.0%

* common
PW20: 31.96% (29.69% - 33.33%)
MAE: 12.645 (12.214 - 13.228)
R^2: 0.073 (0.006 - 0.092)
Time: 103ms (1ms - 857ms)

* clinical
PW20: 36.37% (34.00% - 38.52%)
MAE: 11.367 (10.785 - 12.248)
R^2: 0.161 (0.061 - 0.204)
Time: 484ms (2ms - 3492ms)

* liu
PW20: 41.89% (37.42% - 47.24%)
MAE: 9.990 (9.289 - 11.178)
R^2: 0.254 (0.096 - 0.319)
Time: 648ms (2ms - 3888ms)

3 rows written to 'results/exp2a/iwpc-by-params_2018-09-12_12-38.csv'

EXP 2B: Comparison of Treatments for Missing Data.
Comparing missing-data treatments on 'datasets/Withheld/PATH-V.csv' ...

Treatment: excision, n_train = 3696, n_val = 925
100.0%
Treatment: mode, n_train = 3696, n_val = 925
100.0%
Treatment: mean, n_train = 3696, n_val = 925
100.0%

* excision
PW20: 34.15% (32.22% - 36.11%)
MAE: 11.377 (11.198 - 11.871)
R^2: 0.104 (0.031 - 0.129)
Time: 318ms (1ms - 2267ms)

* mode
PW20: 34.16% (32.22% - 36.43%)
MAE: 11.376 (11.197 - 11.857)
R^2: 0.104 (0.032 - 0.130)
Time: 316ms (1ms - 2279ms)

* mean
PW20: 34.15% (32.22% - 36.32%)
MAE: 11.377 (11.197 - 11.856)
R^2: 0.104 (0.032 - 0.130)
Time: 316ms (1ms - 2237ms)

3 rows written to 'results/exp2b/path-by-treatment_2018-09-12_12-38.csv'
Comparing missing-data treatments on 'datasets/Withheld/IWPC-V.csv' ...

Treatment: excision, n_train = 1557, n_val = 906
100.0%
Treatment: mode, n_train = 3623, n_val = 906
100.0%
Treatment: mean, n_train = 3623, n_val = 906
100.0%

* excision
PW20: 40.08% (33.11% - 45.58%)
MAE: 10.422 (9.423 - 12.346)
R^2: 0.242 (0.034 - 0.314)
Time: 283ms (1ms - 1852ms)

* mode
PW20: 41.80% (37.42% - 46.80%)
MAE: 10.031 (9.309 - 11.301)
R^2: 0.250 (0.084 - 0.322)
Time: 657ms (2ms - 3837ms)

* mean
PW20: 41.83% (37.42% - 46.47%)
MAE: 9.990 (9.298 - 11.175)
R^2: 0.254 (0.097 - 0.314)

```

Time: 646ms (2ms - 3833ms)

3 rows written to 'results/exp2b/iwpc-by-treatment_2018-09-12_12-38.csv'

EXP 3: Novel Algorithms for Warfarin Dosing.

Loaded 'datasets/IWPC-T.csv': 3623/3623 entries used, 58 features

Loaded 'datasets/Withheld/IWPC-V.csv': 906/906 entries used, 58 features

Validating 13 models: validate-iwpc-liu-mean

100.0%

* dLR

PW20: 42.38% (42.38% - 42.38%)

MAE: 9.636 (9.636 - 9.636)

R²: 0.302 (0.302 - 0.302)

Time: 11ms (11ms - 12ms)

* GG

PW20: 20.39% (19.32% - 23.29%)

MAE: 19.122 (18.468 - 19.423)

R²: 0.000 (0.000 - 0.000)

Time: 0ms (0ms - 0ms)

* RT

PW20: 40.95% (40.95% - 40.95%)

MAE: 10.387 (10.387 - 10.387)

R²: 0.231 (0.231 - 0.231)

Time: 6ms (6ms - 6ms)

* MLP

PW20: 45.28% (42.27% - 48.90%)

MAE: 9.489 (9.140 - 9.823)

R²: 0.301 (0.282 - 0.318)

Time: 4601ms (2177ms - 6491ms)

* SVR

PW20: 45.14% (44.59% - 45.36%)

MAE: 9.385 (9.379 - 9.390)

R²: 0.294 (0.294 - 0.295)

Time: 48ms (45ms - 52ms)

* LARS

PW20: 43.27% (43.27% - 43.27%)

MAE: 9.595 (9.595 - 9.595)

R²: 0.310 (0.310 - 0.310)

Time: 230ms (227ms - 235ms)

* PL1

PW20: 44.48% (44.48% - 44.48%)

MAE: 9.572 (9.572 - 9.572)

R²: 0.310 (0.310 - 0.310)

Time: 234ms (228ms - 236ms)

* PL2

PW20: 46.55% (45.92% - 47.46%)

MAE: 9.342 (9.316 - 9.359)

R²: 0.325 (0.322 - 0.328)

Time: 741ms (689ms - 889ms)

* PL3

PW20: 46.60% (45.25% - 47.57%)

MAE: 9.190 (9.141 - 9.266)

R²: 0.311 (0.305 - 0.315)

Time: 418ms (407ms - 424ms)

* PL4

PW20: 46.74% (46.03% - 47.35%)

MAE: 9.338 (9.315 - 9.365)

R²: 0.326 (0.322 - 0.328)

Time: 710ms (692ms - 745ms)

* PL5

PW20: 46.92% (45.81% - 47.68%)

MAE: 9.184 (9.125 - 9.234)

R²: 0.313 (0.308 - 0.320)

Time: 480ms (468ms - 494ms)

* PL6

PW20: 44.38% (44.04% - 44.48%)

MAE: 9.446 (9.443 - 9.453)

R²: 0.315 (0.314 - 0.315)

Time: 622ms (608ms - 633ms)

* PL7
PW20: 43.27% (43.27% - 43.27%)
MAE: 9.606 (9.586 - 9.636)
R^2: 0.316 (0.314 - 0.317)
Time: 186ms (169ms - 208ms)

13 rows written to 'results/exp3/iwpc-by-alg_2018-09-12_12-38_val.csv'
Loaded 'datasets/RAW/IWPC/IWPC-whole-filtered.csv': 3894/3894 entries used, 58 features
Validating 13 models: validate-iwpc-liu-mean-mccv
CV technique: mccv
Comparing 13 models for 100 iterations on validate-iwpc-liu-mean-mccv ...
N (test): 779 x 100 repeats

* dLR
PW20: 46.64% (43.13% - 50.83%)
MAE: 8.790 (7.896 - 12.970)
R^2: 0.433 (0.000 - 0.538)
Time: 10ms (9ms - 15ms)

* GG
PW20: 20.03% (16.56% - 23.49%)
MAE: 19.384 (17.748 - 20.976)
R^2: 0.000 (0.000 - 0.000)
Time: 0ms (0ms - 0ms)

* RT
PW20: 42.74% (38.38% - 46.73%)
MAE: 9.697 (8.674 - 10.958)
R^2: 0.348 (0.206 - 0.442)
Time: 6ms (5ms - 8ms)

* MLP
PW20: 44.27% (37.48% - 50.83%)
MAE: 10.140 (7.953 - 51.657)
R^2: 0.333 (0.000 - 0.497)
Time: 6185ms (2354ms - 10697ms)

* SVR
PW20: 47.84% (42.75% - 53.27%)
MAE: 8.674 (7.711 - 12.773)
R^2: 0.420 (0.000 - 0.525)
Time: 34ms (25ms - 43ms)

* LARS
PW20: 47.20% (42.49% - 50.96%)
MAE: 8.778 (7.874 - 11.569)
R^2: 0.429 (0.000 - 0.526)
Time: 223ms (186ms - 273ms)

* PL1
PW20: 46.87% (43.26% - 50.96%)
MAE: 8.750 (7.901 - 12.887)
R^2: 0.448 (0.000 - 0.540)
Time: 210ms (193ms - 253ms)

* PL2
PW20: 47.98% (44.80% - 52.12%)
MAE: 8.509 (7.687 - 9.113)
R^2: 0.472 (0.322 - 0.555)
Time: 663ms (619ms - 996ms)

* PL3
PW20: 48.78% (44.54% - 54.04%)
MAE: 8.420 (7.567 - 9.167)
R^2: 0.463 (0.323 - 0.547)
Time: 390ms (348ms - 501ms)

* PL4
PW20: 47.97% (44.29% - 51.86%)
MAE: 8.513 (7.675 - 9.140)
R^2: 0.472 (0.321 - 0.559)
Time: 650ms (609ms - 854ms)

* PL5
PW20: 48.81% (44.29% - 53.15%)
MAE: 8.413 (7.524 - 9.181)
R^2: 0.464 (0.331 - 0.544)
Time: 429ms (394ms - 468ms)

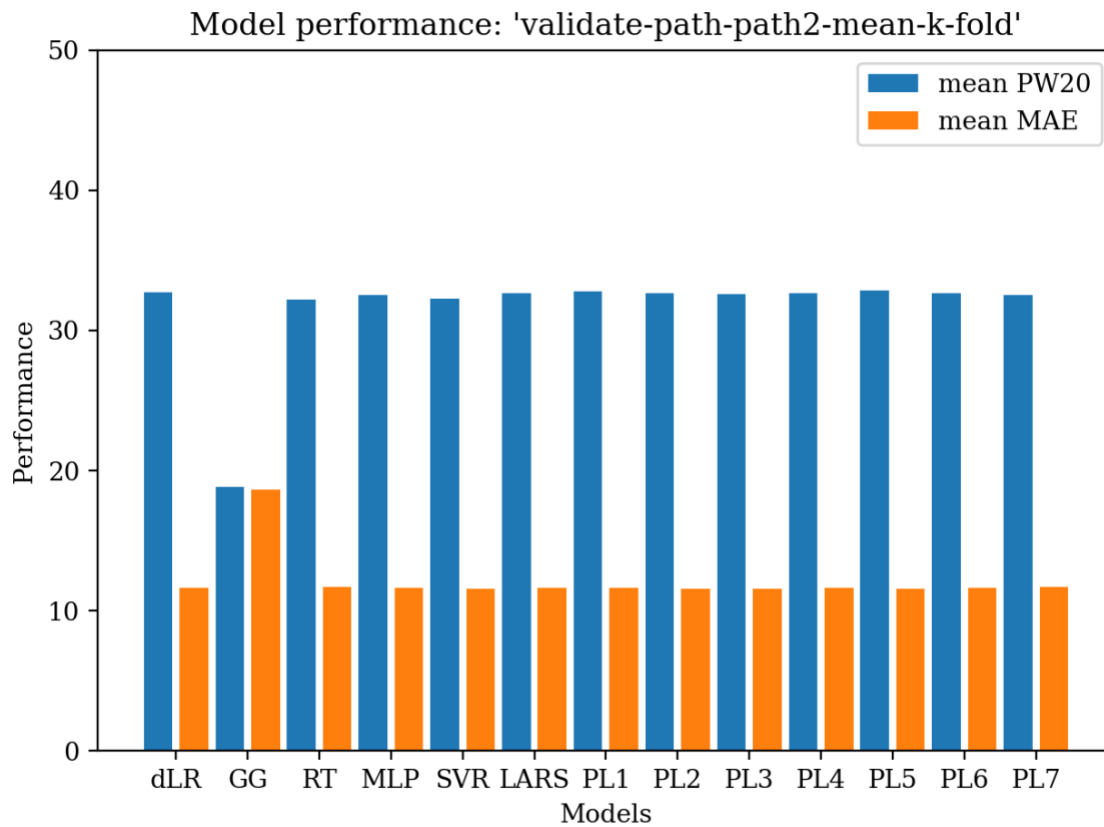
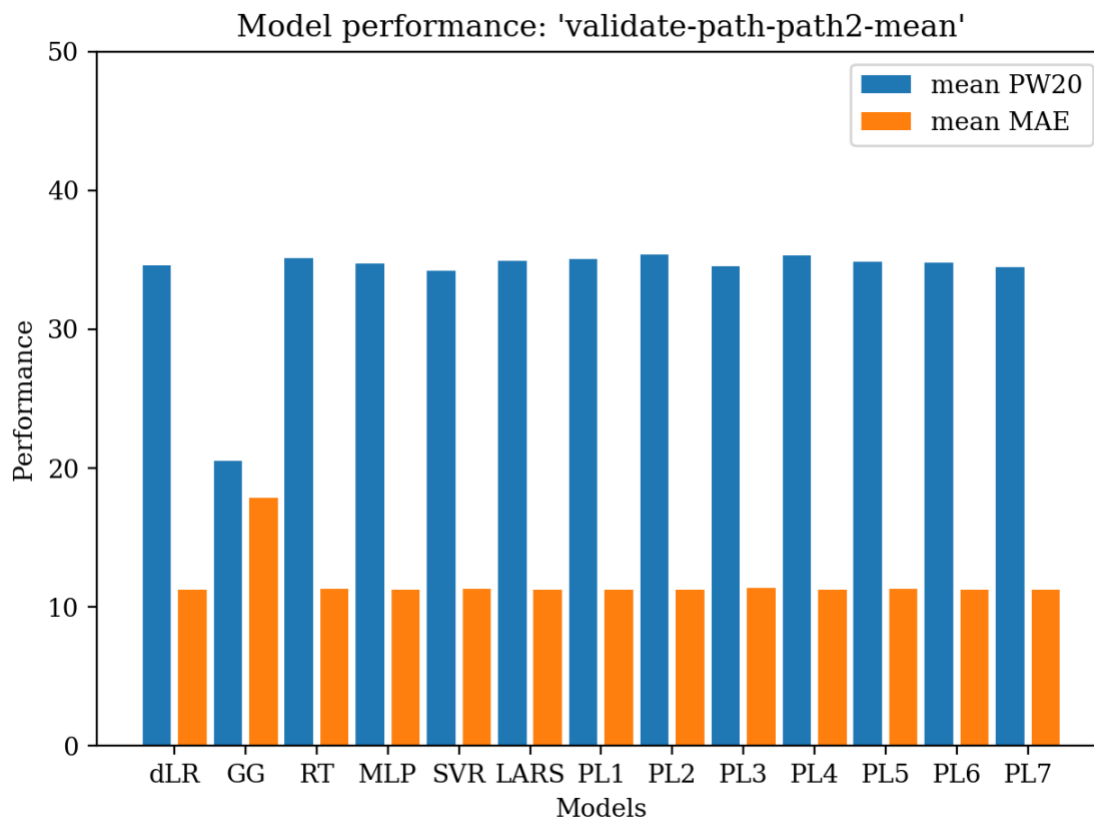

```
* PL6
PW20: 47.56% (43.77% - 51.86%)
MAE: 8.576 (7.750 - 9.297)
R^2: 0.454 (0.251 - 0.550)
Time: 546ms (508ms - 728ms)
```

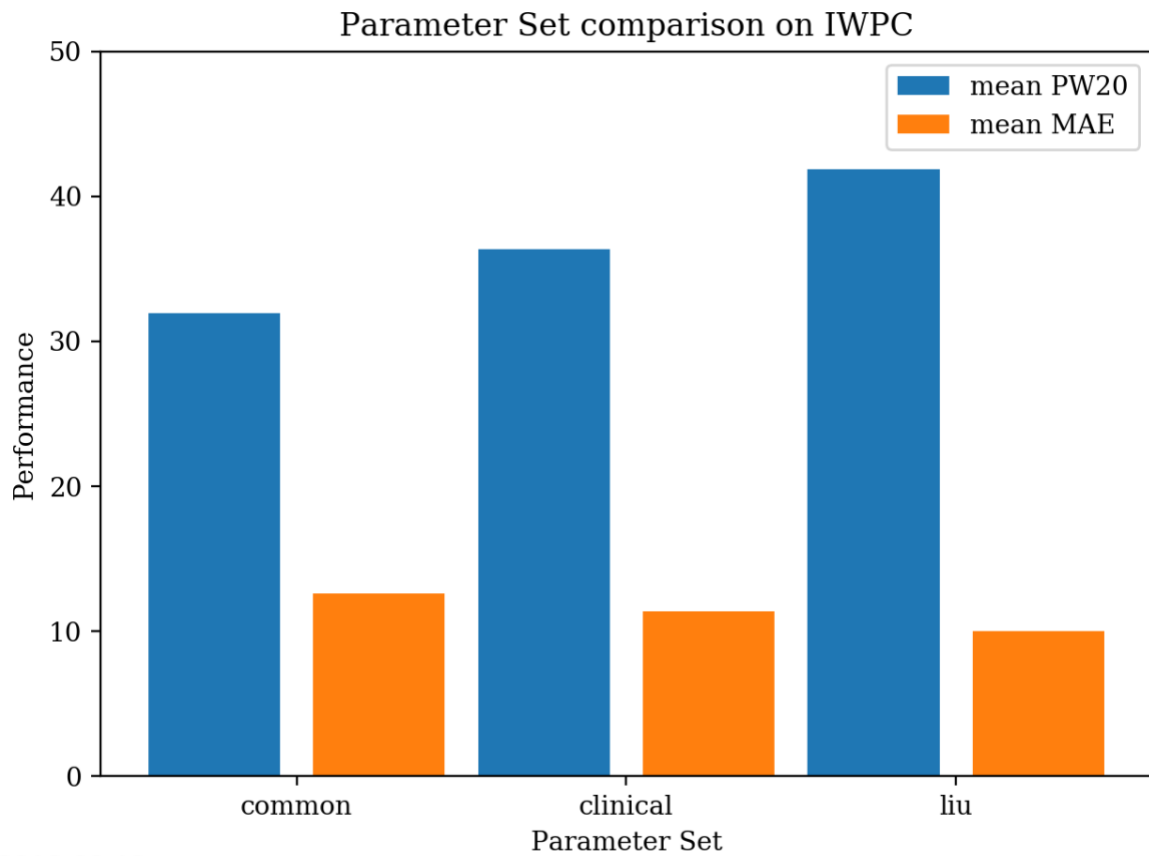
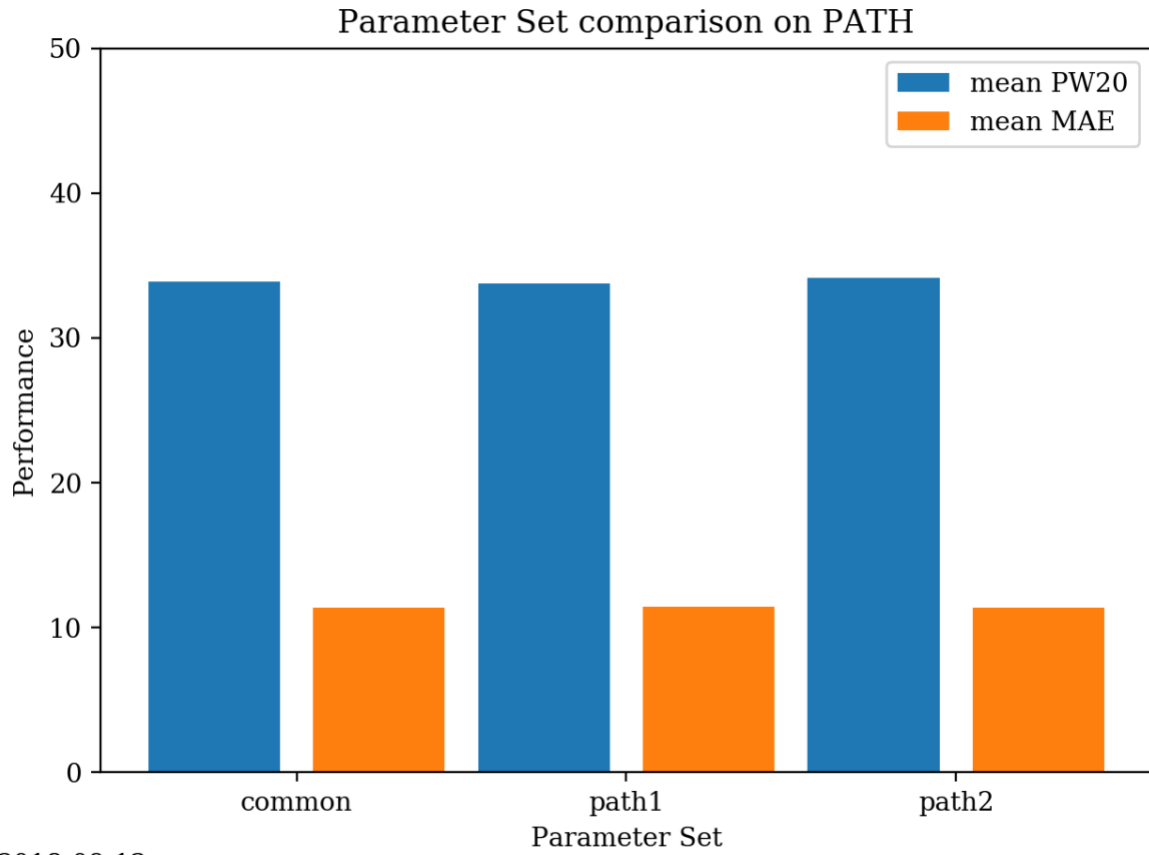
```
* PL7
PW20: 45.22% (2.05% - 50.83%)
MAE: 9.486 (7.992 - 63.832)
R^2: 0.429 (0.000 - 0.527)
Time: 147ms (116ms - 202ms)
```

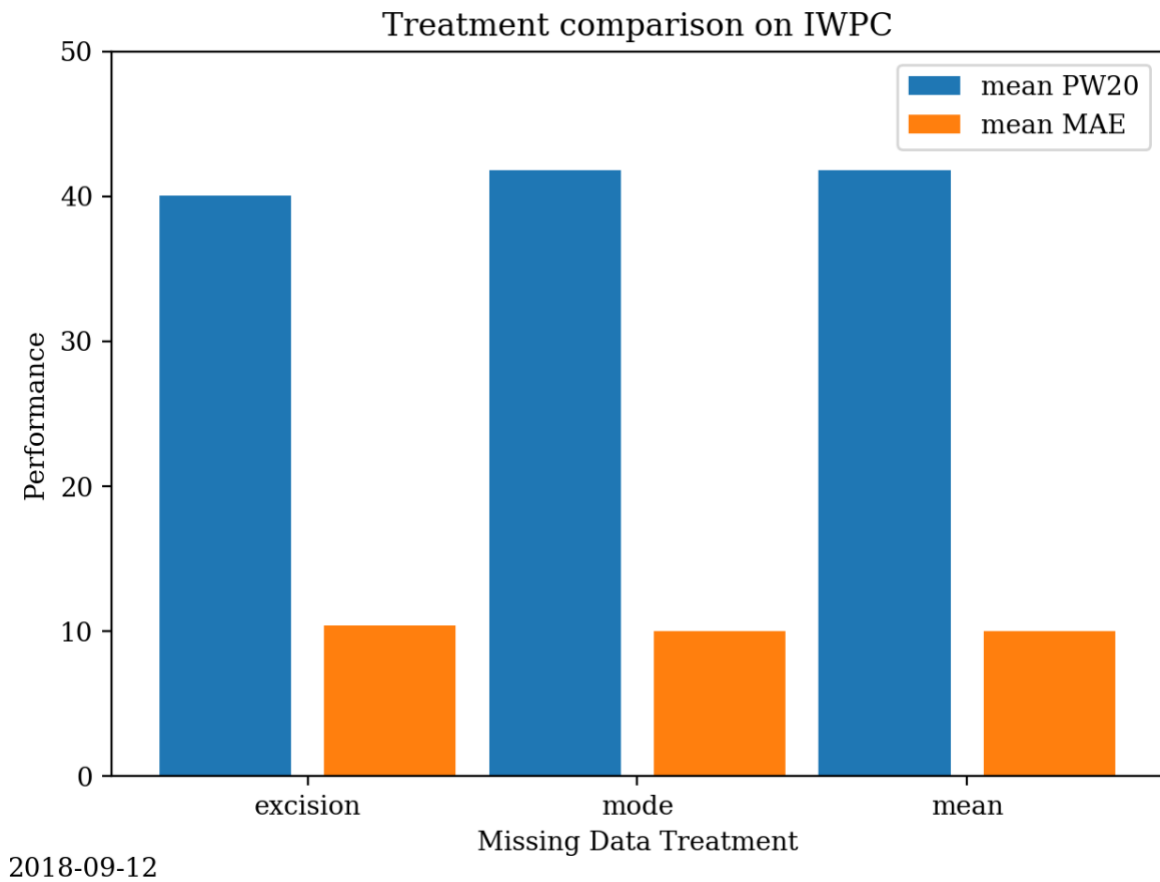
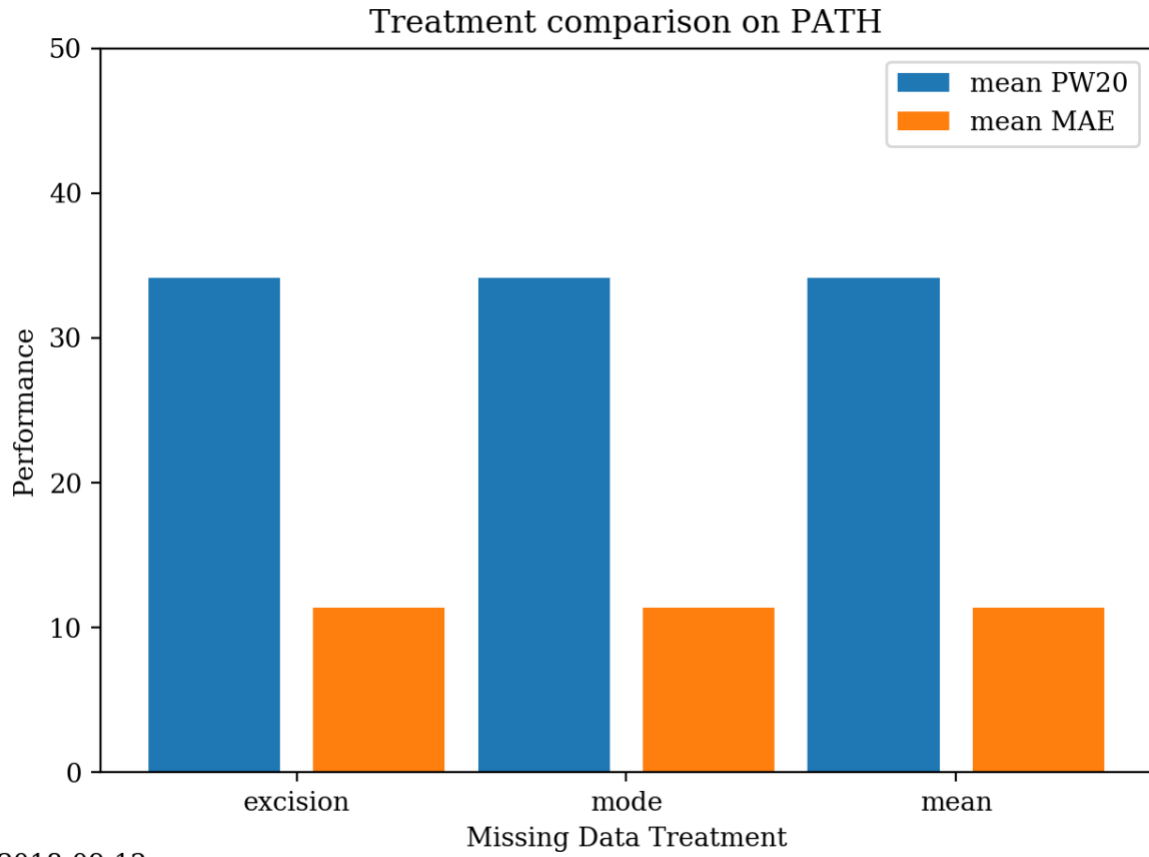
```
13 rows written to 'results/exp3/iwpc-by-alg_2018-09-12_12-38_xval.csv'
```

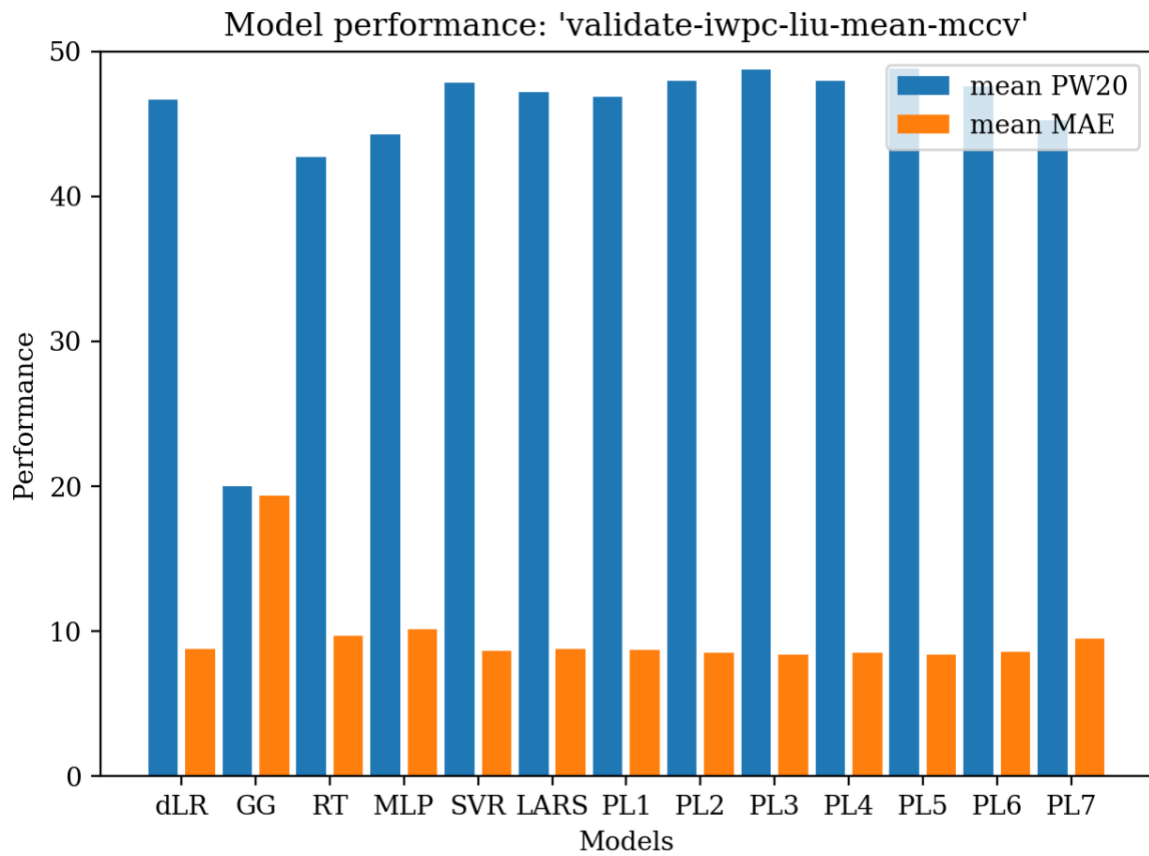
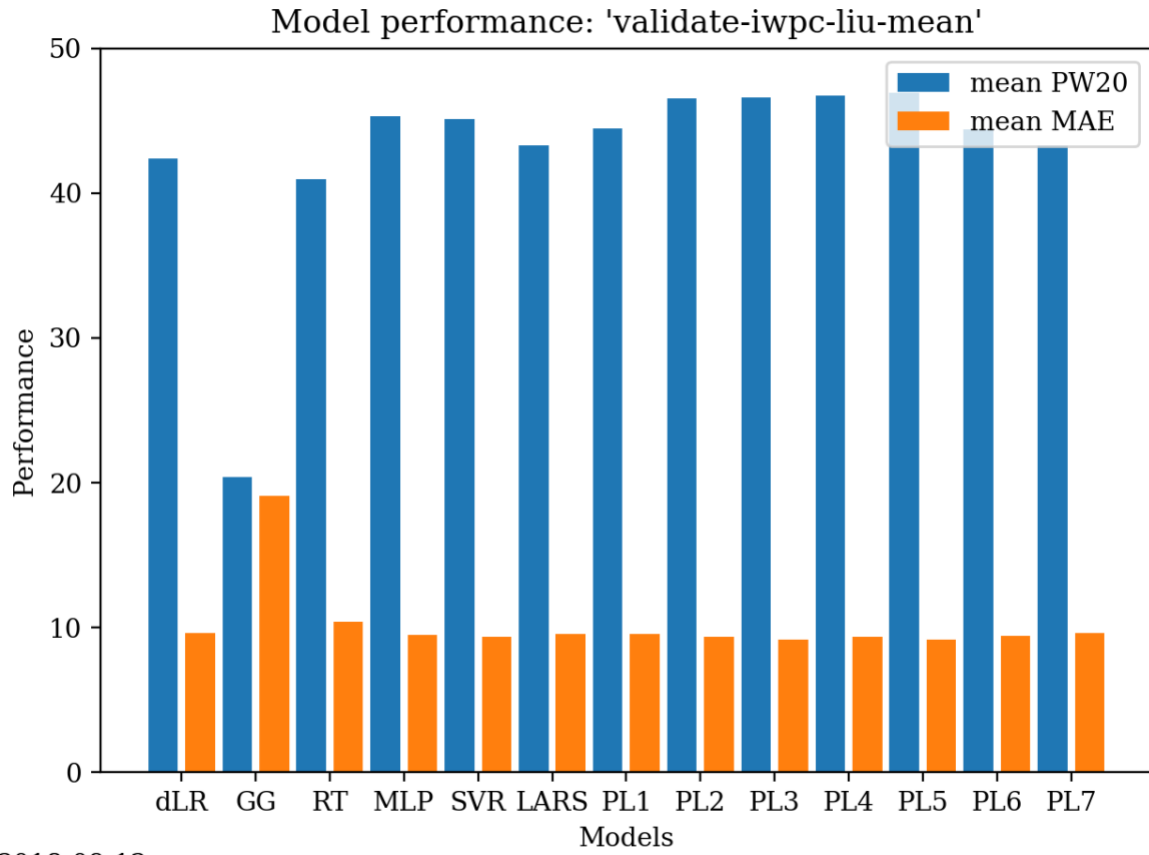
```
gianlucatruda at gianluca in ~/warfit on master [$]
$
```

Additional Generated Graphics

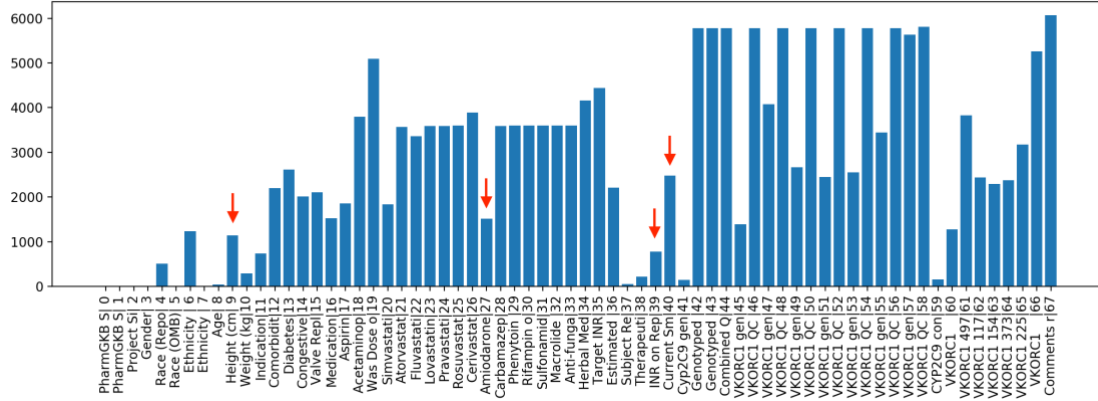




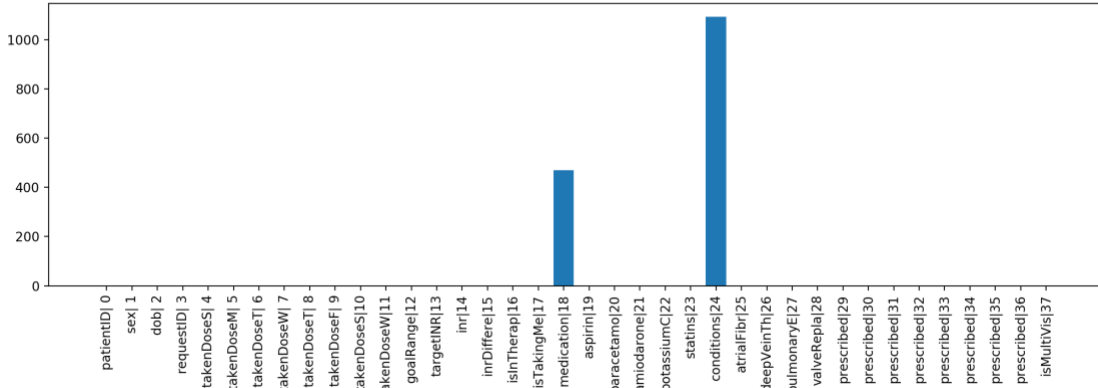




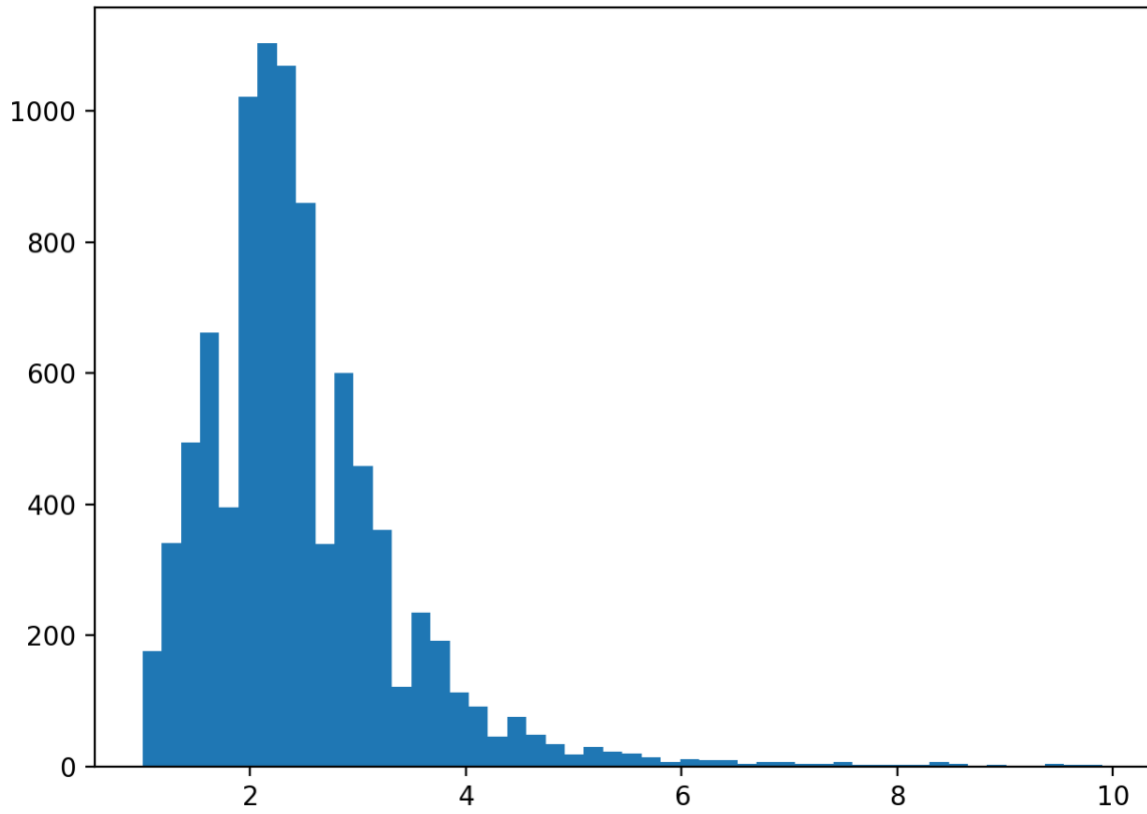
Frequency of missing values: ../datasets/iwpc_ethnicity_raw.csv



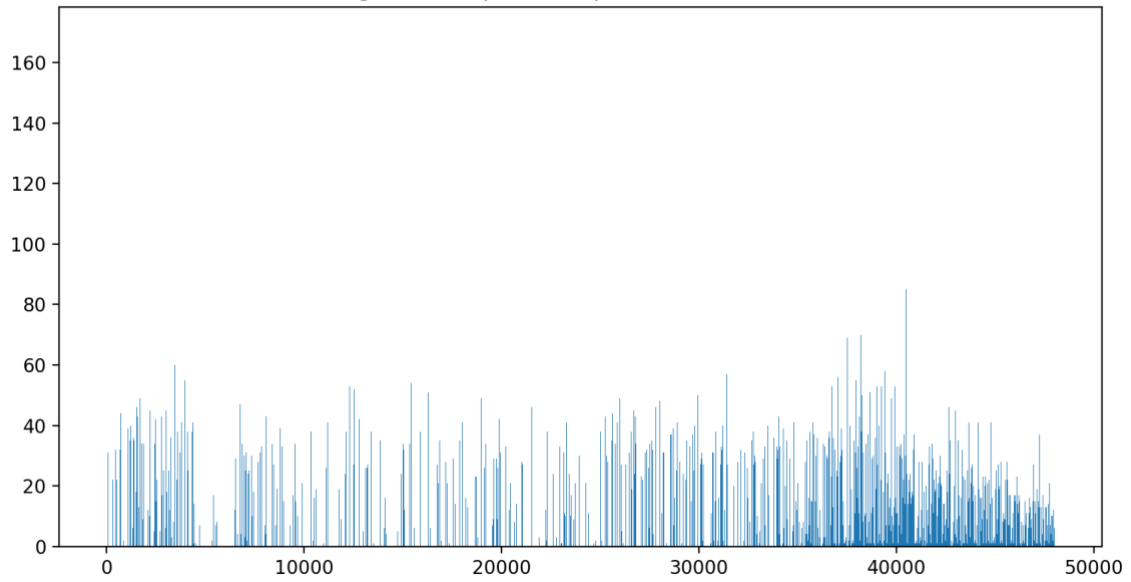
Frequency of missing values: datasets/PATH-T.csv



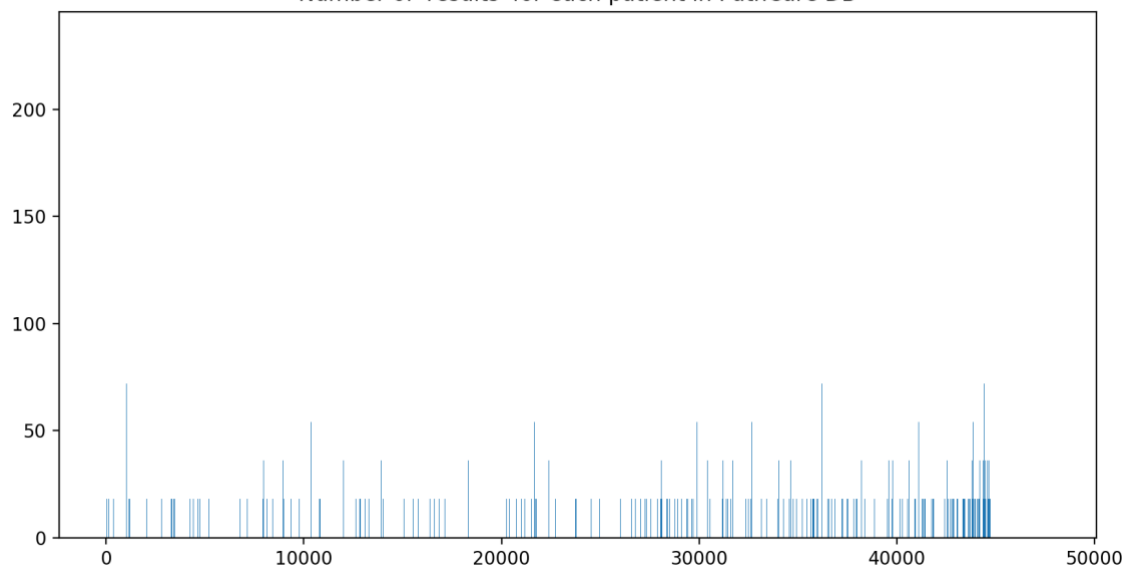
Histogram of valid INR values extracted from PathCare DB



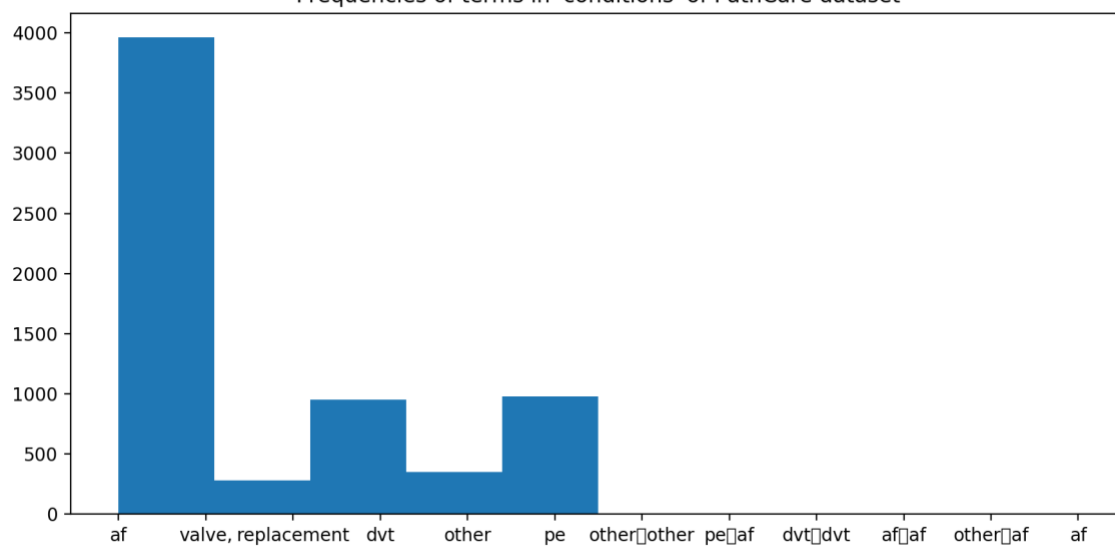
Histogram of requests for patients in PathCare DB



Number of 'results' for each patient in PathCare DB



Frequencies of terms in 'conditions' of PathCare dataset



Dataset Distribution Analyses

Analysis of ../datasets/iwpc_ethnicity_raw.csv

count 6256
unique 6256
top PA152031497
freq 1
Name: PharmGKB Subject ID, dtype: object

count 6256
unique 6256
top PA135312663
freq 1
Name: PharmGKB Sample ID, dtype: object

count 6256.000000
mean 9.425032
std 6.633886
min 1.000000
25% 3.000000
50% 8.000000
75% 15.000000
max 22.000000
Name: Project Site, dtype: float64

count 6252
unique 2
top male
freq 3569
Name: Gender, dtype: object

count 5750
unique 28
top White
freq 2539
Name: Race (Reported), dtype: object

count 6256
unique 4
top White
freq 3425
Name: Race (OMB), dtype: object

count 5017
unique 28
top not Hispanic or Latino
freq 3563
Name: Ethnicity (Reported), dtype: object

count 6256
unique 4
top not Hispanic or Latino
freq 4747
Name: Ethnicity (OMB), dtype: object

count 6214
unique 9
top 70 - 79
freq 1676
Name: Age, dtype: object

count 5110.000000
mean 168.318322
std 10.856168
min 124.970000
25% 160.020000
50% 167.890000
75% 176.530000
max 202.000000
Name: Height (cm), dtype: float64

count 5969.000000
mean 78.735652
std 22.224552
min 30.000000
25% 63.000000
50% 76.000000
75% 90.700000
max 237.700000
Name: Weight (kg), dtype: float64

count 5515
unique 53
top 3
freq 1858
Name: Indication for Warfarin Treatment, dtype: object

count 4057
unique 1632
top No comorbidities
freq 280
Name: Comorbidities, dtype: object

count 3638.000000
mean 0.179219
std 0.383589
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000
Name: Diabetes, dtype: float64

count 4244.000000
mean 0.180254
std 0.384445
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Congestive Heart Failure and/or Cardiomyopathy, dtype: float64

count 4147.000000
mean 0.251266
std 0.433793
min 0.000000
25% 0.000000
50% 0.000000
75% 1.000000
max 1.000000

Name: Valve Replacement, dtype: float64

count 4732
unique 2188
top not aspirin; not simvastatin; not amiodarone
freq 460

Name: Medications, dtype: object

count 4396.000000
mean 0.244995
std 0.430133
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Aspirin, dtype: float64

count 2463.000000
mean 0.222493
std 0.416005
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Acetaminophen or Paracetamol (Tylenol), dtype: float64

count 1159.000000
mean 0.006040
std 0.077514
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Was Dose of Acetaminophen or Paracetamol (Tylenol) >1300mg/day, dtype: float64

count 4417
unique 4
top 0
freq 3695

Name: Simvastatin (Zocor), dtype: object

count 2687.000000
mean 0.095274
std 0.293647

min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Atorvastatin (Lipitor), dtype: float64

count 2895.000000
mean 0.004491
std 0.066872
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Fluvastatin (Lescol), dtype: float64

count 2663.000000
mean 0.015021
std 0.121658
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Lovastatin (Mevacor), dtype: float64

count 2665.000000
mean 0.027017
std 0.162163
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Pravastatin (Pravachol), dtype: float64

count 2656.000000
mean 0.006401
std 0.079762
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Rosuvastatin (Crestor), dtype: float64

count 2368.0
mean 0.0
std 0.0
min 0.0
25% 0.0
50% 0.0
75% 0.0
max 0.0

Name: Cerivastatin (Baycol), dtype: float64

count 4738.000000
mean 0.064795
std 0.246190
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Amiodarone (Cordarone), dtype: float64

count 2662.000000
mean 0.012021
std 0.109000
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Carbamazepine (Tegretol), dtype: float64

count 2656.000000
mean 0.009789
std 0.098473
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Phenytoin (Dilantin), dtype: float64

count 2655.000000
mean 0.001507
std 0.038793
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Rifampin or Rifampicin, dtype: float64

count 2654.000000
mean 0.006405
std 0.079792
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Sulfonamide Antibiotics, dtype: float64

count 2655.000000
mean 0.002637
std 0.051289
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000

max 1.000000
Name: Macrolide Antibiotics, dtype: float64

count 2658.000000
mean 0.006396
std 0.079733
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Anti-fungal Azoles, dtype: float64

count 2097.000000
mean 0.488794
std 0.499994
min 0.000000
25% 0.000000
50% 0.000000
75% 1.000000
max 1.000000

Name: Herbal Medications, Vitamins, Supplements, dtype: float64

count 1815.000000
mean 2.508567
std 0.210978
min 1.300000
25% 2.500000
50% 2.500000
75% 2.500000
max 3.500000

Name: Target INR, dtype: float64

count 4050
unique 13
top 2-3
freq 2846

Name: Estimated Target INR Range Based on Indication, dtype: object

count 6207.000000
mean 0.956017
std 0.205073
min 0.000000
25% 1.000000
50% 1.000000
75% 1.000000
max 1.000000

Name: Subject Reached Stable Dose of Warfarin, dtype: float64

count 6037.000000
mean 31.710537
std 17.113309
min 2.100000
25% 20.000000
50% 28.000000
75% 40.000000
max 315.000000

Name: Therapeutic Dose of Warfarin, dtype: float64

count 5475.000000
mean 2.365697
std 0.463113
min 0.800000
25% 2.100000
50% 2.400000
75% 2.610000
max 6.100000

Name: INR on Reported Therapeutic Dose of Warfarin, dtype: float64

count 3776.000000
mean 0.138506
std 0.345476
min 0.000000
25% 0.000000
50% 0.000000
75% 0.000000
max 1.000000

Name: Current Smoker, dtype: float64

count 6110
unique 11
top *1/*1
freq 4565

Name: Cyp2C9 genotypes, dtype: object

count 478
unique 3
top *1/*1
freq 414

Name: Genotyped QC Cyp2C9*2, dtype: object

count 478
unique 3
top *1/*1
freq 430

Name: Genotyped QC Cyp2C9*3, dtype: object

count 478
unique 6
top *1/*1
freq 370

Name: Combined QC CYP2C9, dtype: object

count 4865
unique 3
top A/G
freq 1727

Name: VKORC1 genotype: -1639 G>A (3673); chr16:31015190; rs9923231; C/T, dtype: object

count 478
unique 3
top G/G
freq 206

Name: VKORC1 QC genotype: -1639 G>A (3673); chr16:31015190; rs9923231; C/T, dtype: object

count 2179
unique 3
top T/T
freq 1506
Name: VKORC1 genotype: 497T>G (5808); chr16:31013055; rs2884737; A/C, dtype: object

count 478
unique 3
top T/T
freq 329
Name: VKORC1 QC genotype: 497T>G (5808); chr16:31013055; rs2884737; A/C, dtype: object

count 3592
unique 3
top T/T
freq 1525
Name: VKORC1 genotype: 1173 C>T(6484); chr16:31012379; rs9934438; A/G, dtype: object

count 478
unique 3
top C/C
freq 206
Name: VKORC1 QC genotype: 1173 C>T(6484); chr16:31012379; rs9934438; A/G, dtype: object

count 3807
unique 3
top C/G
freq 1368
Name: VKORC1 genotype: 1542G>C (6853); chr16:31012010; rs8050894; C/G, dtype: object

count 478
unique 3
top C/G
freq 193
Name: VKORC1 QC genotype: 1542G>C (6853); chr16:31012010; rs8050894; C/G, dtype: object

count 3702
unique 3
top G/G
freq 1990
Name: VKORC1 genotype: 3730 G>A (9041); chr16:31009822; rs7294; A/G, dtype: object

count 478
unique 3
top G/G
freq 215
Name: VKORC1 QC genotype: 3730 G>A (9041); chr16:31009822; rs7294; A/G, dtype: object

count 2809
unique 3
top T/T
freq 1218
Name: VKORC1 genotype: 2255C>T (7566); chr16:31011297; rs2359612; A/G, dtype: object

count 478
unique 3

top C/C
freq 185
Name: VKORC1 QC genotype: 2255C>T (7566); chr16:31011297; rs2359612; A/G, dtype: object

count 627
unique 3
top C/C
freq 408
Name: VKORC1 genotype: -4451 C>A (861); Chr16:31018002; rs17880887; A/C, dtype: object

count 451
unique 3
top C/C
freq 272
Name: VKORC1 QC genotype: -4451 C>A (861); Chr16:31018002; rs17880887; A/C, dtype: object

count 6099
unique 11
top *1/*1
freq 4567
Name: CYP2C9 consensus, dtype: object

count 4983
unique 3
top A/G
freq 1764
Name: VKORC1 -1639 consensus, dtype: object

count 2423
unique 3
top T/T
freq 1661
Name: VKORC1 497 consensus, dtype: object

count 3820
unique 3
top T/T
freq 1556
Name: VKORC1 1173 consensus, dtype: object

count 3960
unique 3
top C/G
freq 1434
Name: VKORC1 1542 consensus, dtype: object

count 3882
unique 3
top G/G
freq 2056
Name: VKORC1 3730 consensus, dtype: object

count 3080
unique 3
top T/T
freq 1273
Name: VKORC1 2255 consensus, dtype: object

count 998
unique 3
top C/C
freq 629
Name: VKORC1 -4451 consensus, dtype: object

count 398
unique 61
top
freq 211
Name: Comments regarding Project Site Dataset, dtype: object

Analysis of .././Pathcare data/pathcare_preliminary.csv

count 8983.000000
mean 31079.275186
std 14853.386348
min 23.000000
25% 19901.500000
50% 35426.000000
75% 44628.000000
max 49275.000000
Name: patientID, dtype: float64

count 8983
unique 2
top F
freq 4867
Name: sex, dtype: object

count 8983.000000
mean 1948.110653
std 14.341502
min 1917.000000
25% 1938.000000
50% 1946.000000
75% 1957.000000
max 2004.000000
Name: dob, dtype: float64

count 8.983000e+03
mean 1.134186e+06
std 2.627328e+03
min 1.129636e+06
25% 1.131914e+06
50% 1.134183e+06
75% 1.136458e+06
max 1.138746e+06
Name: requestID, dtype: float64

count 8983.000000
mean 4.062501
std 2.714893
min 0.000000
25% 2.500000
50% 5.000000
75% 5.000000
max 52.500000
Name: takenDoseSun, dtype: float64

count 8983.000000
mean 4.158163
std 2.695901
min 0.000000
25% 2.500000
50% 5.000000
75% 5.000000
max 30.000000

Name: takenDoseMon, dtype: float64

count 8983.000000
mean 4.118478
std 2.786863
min 0.000000
25% 2.500000
50% 5.000000
75% 5.000000
max 75.000000

Name: takenDoseTue, dtype: float64

count 8983.000000
mean 4.139377
std 2.713027
min 0.000000
25% 2.500000
50% 5.000000
75% 5.000000
max 35.500000

Name: takenDoseWed, dtype: float64

count 8983.000000
mean 4.099309
std 2.699505
min 0.000000
25% 2.500000
50% 5.000000
75% 5.000000
max 50.000000

Name: takenDoseThu, dtype: float64

count 8983.000000
mean 4.142992
std 2.853814
min 0.000000
25% 2.500000
50% 5.000000
75% 5.000000
max 75.000000

Name: takenDoseFri, dtype: float64

count 8983.000000
mean 4.098891
std 2.657331
min 0.000000
25% 2.500000
50% 5.000000
75% 5.000000
max 30.000000

Name: takenDoseSat, dtype: float64

count 8983.000000
mean 28.819712
std 16.713059
min 0.000000
25% 17.500000

50% 27.500000
75% 37.500000
max 210.000000
Name: takenDoseWeek, dtype: float64

count 8983
unique 43
top 2.0-3.0
freq 7902
Name: goalRange, dtype: object

count 8983.000000
mean 2.473506
std 0.940326
min 1.000000
25% 1.900000
50% 2.300000
75% 2.800000
max 9.900000
Name: inr, dtype: float64

count 8983.000000
mean 0.663475
std 0.472547
min 0.000000
25% 0.000000
50% 1.000000
75% 1.000000
max 1.000000
Name: isTakingMedication, dtype: float64

count 5960
unique 4599
top as per previous form
freq 62
Name: medications, dtype: object

count 6263
unique 5
top af
freq 3698
Name: conditions, dtype: object

count 8983.000000
mean 4.178504
std 2.614470
min 0.000000
25% 2.500000
50% 5.000000
75% 5.000000
max 30.000000
Name: prescribedDoseSun, dtype: float64

count 8983.000000
mean 4.302043
std 2.645676
min 0.000000

25% 2.500000
50% 5.000000
75% 5.000000
max 30.000000
Name: prescribedDoseMon, dtype: float64

count 8983.000000
mean 4.239285
std 2.658529
min 0.000000
25% 2.500000
50% 5.000000
75% 5.000000
max 30.000000
Name: prescribedDoseTue, dtype: float64

count 8983.000000
mean 4.270205
std 2.676293
min 0.000000
25% 2.500000
50% 5.000000
75% 5.000000
max 30.000000
Name: prescribedDoseWed, dtype: float64

count 8983.000000
mean 4.192976
std 2.616568
min 0.000000
25% 2.500000
50% 5.000000
75% 5.000000
max 30.000000
Name: prescribedDoseThu, dtype: float64

count 8983.000000
mean 4.269843
std 2.719669
min 0.000000
25% 2.500000
50% 5.000000
75% 5.000000
max 30.000000
Name: prescribedDoseFri, dtype: float64

count 8983.000000
mean 4.238784
std 2.588326
min 0.000000
25% 2.500000
50% 5.000000
75% 5.000000
max 30.000000
Name: prescribedDoseSat, dtype: float64

count 8983.000000

```
mean    29.691640
std     16.377272
min      0.000000
25%     17.500000
50%     27.500000
75%     37.500000
max     210.000000
Name: prescribedDoseWeek, dtype: float64
```