

1 Data-Driven Strategies for *Carbimazole* Titration:
2 Exploring Machine Learning Solutions in
3 Thyrotoxicity Control

4 Reich, Thilo^{1*}, Bakirov, Rashid¹, Budka, Dominika¹, Kelly, Derek²
Smith, James¹ Richardson, Tristan^{1,2}, Budka, Marcin¹

¹Bournemouth University

²University Hospitals Dorset

*Corresponding author & reprint requests: Thilo Reich, treich-contact@runbox.com

ORCID: <https://orcid.org/0000-0001-7705-0987>

5 **Funding received from the Higher Education Innovation Fund**

6 **Keywords:** *Carbimazole* , Hyperthyroidism, Dose prediction, Machine
7 learning, Digital health, Mobile app

Abstract

University Hospitals Dorset (UHD) has over 1,000 thyroid patient contacts annually. These are primarily patients with autoimmune hyperthyroidism treated with *Carbimazole* titration. Dose adjustments are made by a healthcare professional (HCP) based on the results of thyroid function tests. Based on the test results, the HCP prescribes a dose and communicates this to the patient via letter. This is time-consuming and introduces treatment delays. This study aimed to replace some time-intensive manual dose adjustments with a machine learning model to determine *Carbimazole* dosing. This can in the future serve patients with rapid and safe dose determination and ease the pressures on HCPs.

Patient data of 421 hyperthyroidism at UHD were extracted and anonymised. These data were processed and cleaned. A total of 353 patients (83.85%) were included in the study. A wide range of machine learning classification algorithms were tested under different data processing regimes in an iterative approach consisting of an initial model selection followed by a feature selection method to improve performance. All models were evaluated using weighted F1 scores (1=best) and Brier scores to select the best performing model with the highest confidence.

The best performance is achieved using a random forest (RF) approach, resulting in good average F1 scores of 0.731. Based on a balanced assessment considering the accuracy of the prediction ($F1 = 0.755$) and the confidence of the model (Brier score = 0.366), a model was selected with a view to a patient held app. To simulate a use-case, the accumulation of the prediction error over time was assessed. It was determined that an improvement

38 in accuracy is expected if this model was to be deployed in
39 practice.

1 Introduction

Thyroid function plays a crucial role in controlling metabolic rate and cardiovascular function. This regulatory effect is mediated through the thyroid hormones triiodothyronine (T3) and thyroxine (T4). The release of these hormones is, in turn, controlled by the thyroid stimulating hormone (TSH) released from the pituitary gland (1). Any disturbance in the homeostasis of this hormonal system can lead to disease. Generally, thyroid disease can be distinguished between hypothyroidism and hyperthyroidism depending on the level of T3 and T4 production. Hyperthyroidism is a very common endocrine disease with an incidence of 0.2-1.3 % (2) and the most common referral to Endocrinology. In 1977, a UK study reported a prevalence of 2.7% in women and 0.23% in men, highlighting the high disease burden of this disorder nationally (3). There are three avenues of treatment for hyperthyroidism: the use of antithyroid medication, the treatment with radioactive iodine, or surgical removal of the thyroid gland (?). In general, first-time treatment uses antithyroid medications for 6-18 months to allow patients the "opportunity" to naturally gain remission. In the UK, the first line of treatment is *Carbimazole* (?) which involves the titration of the antithyroid drug until the patient reaches an euthyroid state with normal levels of thyroid hormones and TSH. This method is more frequently used and has fewer side effects than block-and-replace approaches (4).

Here, we describe a prototype of a machine learning algorithm to predict an individualised dose of *Carbimazole* based on blood test results. This algorithm could be deployed on a mobile phone

67 application used independently by a patient after an initial clinician-
68 supervised period. Patients are typically reviewed by a clinician every
69 6 weeks and the *Carbimazole* dose is adjusted based on the results
70 of thyroid function tests. Once the test results are available, the
71 clinician usually communicates the adjusted dose to the patient in
72 writing, which is time consuming, costly and introduces delays. The
73 deployment of such a mobile app would be expected to reduce the
74 number of patient contacts by up to 50-70% and would have a notable
75 positive impact on the workload of the clinician and reduce costs for
76 the health care system. In addition patients benefit from patient
77 initiated follow-up (PIFU).

78 1.1 Background

79 The digitisation of healthcare was significantly accelerated by the
80 COVID-19 pandemic, forcing healthcare providers around the world
81 to innovate rapidly (5; 6). This accelerated development will have a
82 long-lasting impact on the way healthcare is delivered. However, it
83 also raises open questions about validation, safety, and accountability,
84 which have already been highlighted pre-pandemic by the EU regu-
85 lations explicitly stating a right to explanation (7). The literature
86 about using machine learning methods to predict drug-dosing regimes
87 for individual patients is mainly focused on cardiovascular medication,
88 thyroidism. In the following paragraphs, a nonexhaustive overview of
89 the literature for each of these sub-areas is described.

90 1.2 Cardiovascular medication

91 Most cardiac-related publications are related to Warfarin dosing, a
92 blood thinner commonly prescribed to patients with atrial fibrillation
93 to reduce the risk of thrombosis.

94 To determine a suitable maintenance dose, a wide range of ap-
95 proaches was found in the literature. These include, among others,
96 machine learning methods such as tree-based predictors (8; 9; 10; 11),
97 k-nearest neighbour (11), support vector machines (12; 10; 9; 8), re-
98 gression such as linear, lasso, or ridge regression (10; 9), random
99 forrest regression (13), ensemble methods (12; 10; 9; 8), boosting
100 algorithms networks (9; 8; 14), and neural networks (12; 10; 9; 14; 15).
101 Additionally, two studies used genetic or evolutionary approaches
102 (16; 17). The main difference between the reviewed publications was
103 the number and variation of the features included to make a prediction.
104 All of the included algorithms used a regression approach to determine
105 the maintenance dose.

107 1.3 Thyroid medication

108 The final and most relevant subject area is the use of predictive algo-
109 rithms for management thyroid disease. In addition to *Carbimazole* ,
110 the drug of interest in this study, other thyroid treatments exist. For
111 example, Levothyroxine is used in replacement therapy after surgical
112 removal, radiation iodine treatment, or hypothyroidism due to other
113 causes. The reviewed studies on levothyroxin specifically tailored
114 their predictions to patients with thyroidectomy. Two study aimed
115 to predict the daily dose requirement using a number of different

116 regression methods (18; 19), while one study predicted whether a dose
117 adjustment was required using a decision tree (DT) (20).

118 No studies were found that predicted antithyroid medication such as
119 *Carbimazole* or propylthiouracil. The literature describing predictive
120 machine learning algorithms for thyroid disease focused extensively
121 on the early detection of thyroid disease (21; 22; 23; 24; 25; 26; 27).
122 Another topic related to thyroidism is the prediction of an appropriate
123 duration of treatment to avoid relapses (28; 29). To the best of the
124 authors' knowledge, there are no published examples of machine
125 learning algorithms used to predict a patient-centered *Carbimazole*
126 dose.

127 2 Methods

128 2.1 Data collection and extraction

129 Data was extracted manually from patient letters, sent after the
130 patient's review. These generally contained the blood test results
131 and the prescribed *Carbimazole* dose. In some cases, additional
132 information about the patient's health or adherence to treatment
133 was also included. Data extraction could not be automated due to
134 IT system limitations and, therefore, was performed manually and
135 entered into a spreadsheet and anonymised. The letters were stored in
136 two separate databases hosted by University Hospitals Dorset (UHD)
137 and accessed using a secure NHS device.

138 2.2 Data processing

139 Most machine learning methods cannot handle missing data without
140 some form of imputation and processing. Our data suffered from
141 overall coverage, making some variables unusable, as they were not
142 recorded often enough and could not be imputed.

143 2.2.1 Data quality

144 A large proportion of the missing data was attributed to human error.
145 This can occur at several stages: copying lab results into a letter,
146 transcription from a dictaphone, or data extraction from the database.
147 Furthermore, patients might not have adhered to prescribed treatment
148 and might have taken medication sporadically or at the wrong dose,
149 adding another human factor that can affect the dataset.

150 2.2.2 Assessing treatment adherence

151 As the data used in this study were extracted from patient letters,
152 in some cases much more information was included than just blood
153 values and prescription dose. Additional data was extracted as free
154 text, including information about side effects or notes that the patient
155 was taking the medication as instructed. Nonadherence to the pre-
156 scribed dose could have implications for predictive algorithms, as it
157 would influence the patient's physiological response and the clinician's
158 decision in a subsequent consultation. Unfortunately, it is difficult to
159 account for these factors in the data, as it is often not clear when a
160 patient changed the medication.

161

162 However, some patients continuously struggled to follow their

163 prescriptions, and these were identified. The most robust way was
 164 based on the information contained in the patient letters, where in
 165 some cases it was mentioned that a patient did not take the prescription
 166 as instructed. This manual approach identified 16 patients (4.9 %).
 167 Furthermore, patients who were prescribed more than 40 mg at any
 168 time were considered non-adherent as this is usually the maximum
 169 dose limit based on UK guidelines (?). In rare cases, clinicians chose
 170 higher doses if they believed that a patient would not regularly take
 171 the medication in the hope that on average an effective dose was
 172 reached. A total of 12 (3.7 %) patients fell into that category. Finally,
 173 the last group of patients where non-concordance was suspected were
 174 patients who were prescribed 40 mg for longer than 90 days with
 175 consistently high T4 levels. This assumes that all patients respond to
 176 *Carbimazole*. In practice, a patient who did not reach normal levels
 177 for a long period of time was reviewed by a clinician. This group
 178 included 3 patients (0.93 %). We concluded that a total of 31 (9.66 %)
 179 patients were not always adhering to the medication regime and hence
 180 excluded them from the dataset A summary is shown in Table ??

181 2.2.3 Determination of prescription dose

182 Most prescription doses were represented as mg/day ensuring that all
 183 prescriptions had one value each day. However, a minority of doses
 184 were prescribed to be taken every second day or n times a week. For
 185 these cases, the average daily dose was calculated. Additionally, some
 186 patients received different doses in the morning and evening. In this
 187 scenario, the daily sum was calculated to allow a single dose as the
 188 prediction target.

189 As *Carbimazole* is not the only antithyroid drug, some patients'
190 treatment regimes were changed to propylthiouracil if side effects
191 occurred. In addition, some patients were prescribed propranolol, a
192 beta-blocker, to treat palpitations, which can occur as a side effect of
193 a hyperthyroidism. In these cases, it was unclear whether this was
194 prescribed along with *Carbimazole* or as a single treatment. For this
195 reason, these instances were removed from the dataset. Additionally,
196 some records were unreadable were removed from the dataset. In
197 total, this affected 7.5 % of the dataset.

198 2.2.4 Binning of doses

199 As clinicians sometimes prescribed doses on alternate days or only
200 a few times per week, several doses were rarely, or in some cases
201 only prescribed once. Such doses were impossible to predict as a
202 classification problem, and, therefore a binning method was used,
203 which assigned such doses to the closest user-defined dose. If a dose
204 fell exactly in the middle of two neighbouring doses, the lower dose
205 was selected. This is because patients tend to prefer lower doses, as
206 these will result in less weight gain (30). The doses used for this
207 binning process were based on the actual prescription pattern and
208 were chosen in consultation with a clinician. The selected binned
209 doses were: 0, 2.5, 5, 10, 15, 20, 40 mg.

210 2.3 Feature engineering

211 Often, the generation of additional features from the available data can
212 improve the performance of predictive algorithms. Several additional
213 features were added to the dataset, which are described below in

214 detail.

215 2.3.1 Dose exponents

216 A feature suggested in the literature was dose exponents, which are
217 calculated by squaring and cubing the previous doses. These have
218 previously been shown to improve the accuracy of the prediction (31).

219 2.3.2 Time between presentations

220 Both the time between presentations and the time since the first
221 presentation in days were calculated based on the contact date. It
222 should be noted that this is not the same as the date the test was
223 performed. However, the date of the blood test was not consistently
224 recorded, so the contact date was used in all cases. The average
225 time between presentations in our datasets was 96 days, roughly
226 corresponding to the desired follow-up interval of three months.

227 2.3.3 Thyroid hormone based features

228 It was suggested by (31) that an important feature is the daily change
229 in T4 level. This was calculated based on the time between pre-
230 sentations and the change in recorded T4 levels. Furthermore, the
231 percentage change of T4 was calculated to give an indication of how
232 fast the change occurred. Additionally, the difference to the normal
233 range was explicitly added, which meant that the approximate mean
234 target (16 pmol/L) was subtracted from the recorded reading. In
235 addition, the percentage change in TSH was calculated to give an
236 indication of how quickly the change occurred.

237 2.3.4 Overt and subclinical thyroidism

238 Clinically, hyperthyroidism is distinguished into overt and subclinical
239 states (?). To ensure that this information was available to the model,
240 Booleans were added for both conditions. For this purpose, we used
241 the following definitions:

- 242 • Overt: $\text{TSH} < 0.5 \text{ mU/L}$ and $\text{T4} > 21 \text{ pmol/L}$
- 243 • Subclinical: $\text{TSH} < 0.5 \text{ mU/L}$ and $10 > \text{T4} < 21 \text{ pmol/L}$

244 2.3.5 T3

245 The levels of T3 are not routinely tested in current practice. This
246 meant that these were not included for each presentation. However,
247 T3 was often recorded once at the beginning of a patient’s treat-
248 ment. Therefore, the ratio of T4:T3 was included as a static patient
249 characteristic, as this can provide additional information on the under-
250 lying pathophysiology of hyperthyroidism (31). Missing values were
251 imputed to the median.

252 2.3.6 Patient exclusion criteria

253 Hyperthyroidism is strongly influenced by pregnancy (32). Our dataset
254 only included 14 (3.3%) patients who experienced one or more pregnan-
255 cies during treatment. These were excluded as they were a minority
256 that could not be adequately represented in any machine learning
257 model. In practice, these patients will have followed a different treat-
258 ment and management pathway. Furthermore, patients with fewer
259 than four presentations were excluded ($n=12$ (2.8%)). Finally, pa-
260 tients who were deemed to not adhere to the treatment regime were

261 removed; see Section 2.2.2 for more details.

262

263 2.3.7 Data exclusion

264 Even though, from a physiological point of view, TSH receptor anti-
265 body levels would be expected to be crucial, these were only available
266 for a minority of patients since blood tests were not readily available
267 at the time. Similarly, blood pressure readings were too infrequent
268 to be viable for inclusion. Finally, weight and BMI were removed as
269 these were also recorded infrequently.

270

271 2.3.8 Imputation

272 Following the data handling procedure described above, most of the
273 features had good coverage. However, the hormone levels of T4 and
274 TSH were missing in some cases (6.8% and 9.7% respectively). As
275 these metrics are physiologically important, a k-Nearest Neighbors
276 (KNN) imputation method was used to approximate the missing values.
277 An indicator was added to identify the imputed values.

278 2.3.9 Manual time adjustments

279 Due to the complexity of the dataset, which included timeseries data
280 for each patient, identifying spurious data was challenging. There-
281 fore, a manual screen was performed to find breaks in the timeline
282 of individual patients. Several of these cases were identified; most
283 were patient contacts in unusually rapid succession. As these mostly
284 followed a thyroid function test with a long follow-up period, it was

285 concluded in consultation with a clinician that this was probably due
286 to a repeat of the test at the next presentation because the follow-
287 up was not timely. In these cases, the initial readings with delayed
288 follow-up were removed because the clinician would have made their
289 decision based on the more recent repeated blood test.

290 2.3.10 Patient grouping

291 Patients could be distinguished into two groups at their first presen-
292 tation based on their clinical state. This was because some patients
293 received hyperthyroidism treatment elsewhere before moving to UHD.
294 Thus, these patients started at a maintenance dose and had well-
295 adjusted hormone levels. This patient group contrasted with patients
296 who presented for the first time at UHD for anti-thyroid treatment.
297 To distinguish between these two groups, hormone levels were evalu-
298 ated on first presentation to our facility. Because the response of T4
299 is faster, the grouping was based on the initial levels of T4. It was
300 assumed that patients with T4 levels >22 pmol/L had not received
301 treatment elsewhere, otherwise it was assumed that the patient had
302 been treated before. The patients were assigned Booleans according
303 to the group they belonged to on their first presentation.

304

305 It was assumed that the time in treatment was an important factor
306 for making an accurate prediction. Therefore, it was expected that the
307 time in treatment for patients who previously received treatment was
308 a confounding factor. To adjust the data to reflect previous treatments
309 for these patients, starting treatment was adjusted to begin at the
310 third presentation where a patient was expected to have T4 levels

311 within the normal range. Furthermore, the average time to the
312 third presentation was used as a starting point. These modifications
313 to the data based on domain knowledge made the prediction task
314 easier as patients at physiologically different stages of treatment were
315 distinguished; see Figure 1.

316 2.3.11 Relapse detection

317 The aim of antithyroid treatment is to normalise thyroid function
318 with the lowest dose of *Carbimazole* with a view to pf a trial cessation
319 of treatment. However, many patients suffer a relapse which requires
320 an adjustment of their *Carbimazole* dose at some point during their
321 treatment time. In this study, we defined a relapse as a TSH level
322 within the normal range (>0.3 mU/L) followed by one below the
323 normal range of 0.3 mU/L. This applied to 61.7% of our patients,
324 which was within the range of the current literature (33). The relapse
325 episodes were labeled with a Boolean. This was important as the
326 treatment of a relapse is different from the initial treatment (Figure
327 2). Part of the reason for this is that relapses were identified faster
328 due to routine tests compared to the initial diagnosis.

329 2.3.12 Polynomial features

330 Additionally, polynomial features were generated, as this approach
331 can improve prediction performance. This generated new features
332 consisting of polynomial combinations of features with a degree less
333 than or equal to the specified degree. In this study, we used a 2-degree
334 polynomial for which the generated features for two features a and b

335 would be:

$$1, a, b, a^2, ab, b^2 \quad (1)$$

336 2.4 Training and testing datasets

337 To prevent contamination of the test set with data from patients in
338 the training set, two approaches were chosen to include patients in
339 the test set. A portion of the patients was chosen, which had not been
340 included in the training set; these were patients who represented the
341 first presentation at the clinic and are referred to as 'new patients'.
342 For the second portion of the patients, each patient's time-series data
343 was split in half, where the first half of presentations was included
344 in the training set and the second half in the testing set referred to
345 as 'previously seen patients'. This represented an easier part of the
346 dataset as all of these patients were on a maintenance dose, which was
347 easier to predict. However, this could affect the representativeness of
348 the two datasets. Ideally, both training and testing datasets should
349 be sampled from the same distribution. This representativeness is
350 difficult to determine as a result of the multidimensionality of our data.

351

352 We used an algorithmic approach to address representativeness
353 in both datasets. Initially, roughly 30% of patients were randomly
354 selected to be included in the test dataset. Of those, 50% were
355 randomly chosen to be included completely in the test set, and for
356 the remaining patients, half of their presentations are included in
357 the training set. This could have created datasets that were skewed
358 towards a certain group of patients. To avoid this, we combined all
359 datasets and chose an equal number of patients from either of the

360 datasets. As a next step, the type of dataset was used as a label and
361 a RF classifier was trained to attempt to predict which data belonged
362 to which dataset. In an ideal scenario, this should have achieved
363 an accuracy equivalent to random chance, as all datasets should be
364 indistinguishable. In many of the randomly chosen data splits, this
365 was not the case, and the classifier achieved relatively high accuracy.
366 To find the most representative data split, this process was repeated,
367 and the dataset, that was the hardest to distinguish, was chosen for all
368 further experiments. This brute-force approach was time consuming,
369 but ensured that patients were not included into certain datasets
370 based on implicit bias.

371 2.5 Balancing of training dataset

372 The dataset is significantly imbalanced because once the patient
373 reached normal hormone levels the maintenance was not changed
374 frequently. Therefore, balancing the prediction classes was necessary
375 to development the predictive algorithms. The balancing strategy
376 was employed in two steps. Initially, all *Carbimazole* classes as de-
377 scribed in section 2.2.4, that contained more entries than the mean
378 number of prescriptions per class were randomly under-sampled to the
379 mean number of samples per class. The remaining classes with fewer
380 entries than the mean were over-sampled using Adaptive Synthetic
381 Sampling Approach for Imbalanced Learning (ADASYN). This used
382 weighted distributions of these minority classes where the weight was
383 the difficulty in making a prediction (34). This method ensured that
384 the dataset was balanced to make predictions of the minority classes
385 possible.

386 2.6 Permutation importance

387 Because not all features in the dataset were necessarily beneficial for
388 a good prediction, feature selection was important. Consequently, we
389 employed a permutation approach, which used a trained algorithm
390 that incorporated all available features. This algorithm was iteratively
391 tested with single feature of the test set randomly permuted. This
392 effectively removed the influence of a particular feature, and the
393 change in the results gave an indication of the importance of the
394 feature (35). The permutation importance scores were then used
395 to iteratively retrain the algorithm with several feature subsets and
396 select the best performing algorithm based on the test F1-score.

397 2.7 The data regimes

398 The data were prepared under several different processing regimes
399 that will be described in the following sections.

400 2.7.1 Presentation adjustment

401 This dataset included the presentation adjustments for the patient
402 groups in the first presentation as described in Section 2.3.10.

403 2.7.2 Relapse separation

404 This dataset included the presentation adjustments for the patient
405 groups at the first presentation as described in Section 2.3.10. Fur-
406 thermore, relapses were detected as described in Section 2.3.11 and
407 treated as separate presentations, which meant that a patient suffering
408 a relapse was considered as starting a new treatment period.

409

410 2.7.3 Gap removal

411 All patients should have had a blood test every 3 months, so longer
412 periods between presentations were separated into individual instances.
413 To allow some leeway, gaps of more than 120 days were considered as
414 a new presentation and were treated as if this patient had been seen
415 for the first time, resetting the time in treatment and the time since
416 the last treatment to 0, thus, representing a new consultation period.

417 2.7.4 Dose correction

418 The prescriptions chosen by the clinician were not always optimal due
419 to human error. As the retrospective dataset allows to see the future
420 response of a patient based on hormone levels at the next presentation,
421 such suboptimal doses were identified and removed.

422

423 The definitions used are as follows:

- 424 • Optimal: a dose that resulted in a future reduction in T4 within
425 the next 120 days.
- 426 • Suboptimal: a dose that resulted in an increase in T4 within the
427 next 120 days.

428 2.8 Algorithm training and evaluation

429 A variety of supervised algorithms were trained on the data in a sys-
430 tematic approach. A gridsearch approach was employed to optimise
431 the best-performing hyperparameters. As a scoring metric for the
432 gridsearch, a weighted Area Under the Receiver Operating Character-
433 istic using the one-vs-one approach (ROC) was used. This computed

the average Area Under Curve (AUC) of all possible pairwise combinations of classes (36). To select the algorithms that performed the best overall, the weighted F1 score measure was used. The algorithm was then subjected to the permutation importance procedure described in Section 2.6. This approach was repeated for all data processing scenarios described in Section 2.7. The initial choice of algorithm was based on the weighted F1 score; however, as this only gives limited insight into the performance, the test sets were further assessed using confusion matrices, the average error in mg, the Brier score, the top 1-3 prediction score, and as ROC. Finally, a logistic regression-based probability calibration was applied to calibrate the algorithm. This allowed a better interpretation of the probability of prediction of the algorithm and could further improve the predictive performance of the model. Although performance improvement was not always achieved, some research suggested that RFs can benefit from a calibration attempt (37).

2.9 Ordinal predictor

An ordinal predictor was fitted because the prediction targets are ordinal in nature and could potentially have benefited from a predictive method tailored to such a problem. For this approach, a binary classifier was fitted for each class. To make a prediction, each individual classifier was used to predict the probability that the sample belonged to each of the specific classes. The class with the highest confidence was chosen as a final prediction. For this approach, the features identified by permutation of the base data set were used. All base classifiers described in Section 2.8 were fitted in an ordinal manner and

460 the best performing version was identified by the F1 cross-validation
461 score.

462 2.10 Confident learner

463 To address the possibility that a label might be incorrect, a confident
464 learner was fitted to the best performing algorithm of the base dataset.
465 This method allowed for an approximation of the likelihood that a label
466 as correct and was reported by (38) to achieve excellent performance
467 in identifying false labels in many different datasets. Thus, a confident
468 learner as described by (38) was used, which estimated the probability
469 that a label was incorrect, allowing for a more robust version of the
470 algorithm to be trained.

471 3 Results

472 The experimental results were divided into six sections, including
473 the four data regimes (the base, gap-removal, dose correction, and
474 polynomial datasets) and the two special predictors (the confident
475 learner, and the ordinal predictors). This is followed by a summary
476 bringing all methods together to give an overview of the best model
477 and data regime. In all scenarios a RF predictor with a decision
478 tree as base estimator performed the best details of each of the data
479 regimes will be discussed below. Data descriptives are shown in the
480 supplementary material .

481 3.0.1 Base dataset

482 In the base data processing regime with minimal adjustments, the RF
483 approach performed the best based on the initial search. This was

484 true for both datasets with and without relapse detection. The overall
485 best performing algorithm was the permuted version with relapse
486 detection with an F1 score of 0.751 and a Brier score of 0.380 for
487 unseen patients Table 3.

488 3.0.2 Gap-removal dataset

489 The second data regime treated patients who had not been seen for
490 more than 120 days as new patients. Again, in this scenario, the RF
491 approach performed best based on the initial search. Similarly to the
492 base dataset, the permutation selected subset of the features resulted
493 in the best performing algorithm. However, this predictor performed
494 worse compared to the predictor trained on the base dataset with an
495 F1 of 0.739 and a Brier score of 0.392 for unseen patients Table 3.

496 3.0.3 Dose-correction dataset

497 When the gap removal approach was combined with the dose correc-
498 tion procedure described in Section 2.7.3, the predictor performance
499 increased. As with the other data regimes, an RF model was selected.
500 This model performed better on unseen patients compared to the
501 algorithm trained in the gap removal data regime. The F1 for unseen
502 patients was 0.743 with a Brier score of 0.382, which means the model
503 was more confident than on the gap removal regime alone. This par-
504 ticular model performed with a F1 of 0.766 and a Brier score of 0.350
505 on the maintenance dataset thus making it the highest performing
506 model in average Table 3.

507 3.0.4 Polynomials

508 The generation of polynomial features for a RF classifier trained in
509 the base data set did not result in better performance compared to
510 the algorithm trained on the conventional feature set. In all scenarios
511 tested, this approach resulted in a reduction in the F1 score compared
512 to the initial approach described in Section 3.0.1. Therefore, this
513 approach was not applied to the other data regimes (Table 3).

514 3.0.5 Confident learner

515 The RF-based confident learner trained on the base model with the
516 permuted feature-set and no relapse detection applied improved the
517 F1 score from 0.693 to 0.696. However, at the same time the Brier
518 score was increased from 0.45 to 0.461, suggesting that the model
519 made less confident predictions. In the case of the maintenance dose,
520 the F1 score was reduced from 0.778 to 0.770 and the Brier score
521 increased slightly from 0.351 to 0.357. In the case of the relapse
522 detection data regime, both models suffered from a reduction in the
523 F1 score and an increase in the Brier score. This suggests that the
524 confident learner approach did not improve the overall performance of
525 the model and, thus, was not applied to other data regimes (Table 3).

526 3.0.6 Ordinal predictor

527 An RF based ordinal predictor was trained and tested for the base
528 data set with the selected features of permutation that showed a
529 general reduction in F1 scores and an increase in Brier score, and
530 was therefore not further developed for the other data regimes. The
531 results are shown in Table 3.

532 3.1 Summary and interpretation of results

533 The experimental results highlight the issue of selecting the "best"
534 model, as several metrics had to be considered. To illustrate overall
535 performance, Figure 3 shows the average performance based on the
536 test dataset comprising new patients, as well as the test set comprising
537 patients only at maintenance doses. For the purpose of the initial
538 model for deployment, a conservative model was chosen. This was
539 based on performance in the new patients, as these are the most
540 difficult samples where the highest accuracy is desired. The model with
541 the highest top-1 score, meaning it achieved the highest percentage
542 of predictions and the lowest Brier score a metric to describe the
543 certainty of the model where the lowest value is the best possible.
544 This model was expected to work best for new patients, but it will
545 also work well for the prediction of maintenance doses. The selected
546 model is the base classifier with relapse detection, but without the
547 removal of suboptimal doses or larger gaps in presentations. The
548 features were reduced by using permutation, but no calibration was
549 applied (Table 3).

550 3.2 Better understanding the model using SHAP values

551 Using SHAP (SHapley Additive exPlanations) values allows the con-
552 tribution of each feature to the final prediction to be quantified using
553 game theory (39). The average SHAP values for each of the features
554 included in the final model are shown in Figure 4. The previous
555 prescribed dose and its derived metrics contribute the most to the
556 final prediction. This was expected as the dose did not change often
557 once a maintenance dose has been reached. Other features of rela-

558 tive high importance are hormone level related features. This was
559 also expected, as the prescribed dose by the physician is based on
560 physiological findings and therefore should be reflected by the model.

561 3.3 Longitudinal error

562 The inclusion of the previously prescribed *Carbimazole* as a feature in
563 the model poses a theoretical problem for the practical use of such a
564 model. Due to the reliance on previous doses and the sequential nature
565 of the prediction task, any prediction error made could carry through
566 the time in treatment for each patient. This means that an error made
567 by the predictor could increase over time if a number of incorrect
568 predictions are made. All metrics reported so far used the clinician-
569 prescribed previous doses. To assess any accumulation of errors over
570 time, a second approach was used in which the prediction made at the
571 previous presentation was used to make subsequent predictions. As
572 the demonstrated algorithm did not aim to replace all patient-clinician
573 contacts but instead to reduce workload, it was assumed that the
574 patient will be seen by a clinician every second presentation, potentially
575 reducing patient contact by 50%. This middle ground serves as a
576 safety net, giving the opportunity to identify any difficulties the user
577 might experience with the new technology, as well as double checking
578 dosing regimes by an experienced clinician to identify any safety issues.

579 The longitudinal assessment showed a more realistic performance
580 over the treatment time of a patient and simulated the final use of
581 the application by a patient. This cumulative error is referred to as a
582 ‘compound error’, in contrast to the conventional error that uses only
583 historical dose prescriptions derived from a clinician. Additionally,

584 a naive predictor using the previously prescribed dose is shown as
 585 a comparison benchmark. A timeline of the change in F1-scores is
 586 illustrated in Figure 5. The F1 score of the compared error initially
 587 was lower than the conventional error, but then soon recovered. The
 588 longer a patient was in treatment, the higher the F1 score, which
 589 was an expected pattern due to fewer dose changes once a patient
 590 had reached a maintenance dose. The general comparison is shown in
 591 Table 1, and the accumulation of errors reduced the F1 score by 9.3 %
 592 and 11.3 % for the previously unseen and previously seen patients,
 593 respectively (Figure 6). The change in prediction accuracy is visualised
 594 in confusion plots where the compound effect can be assessed for each
 595 prediction class (Figure 7). These findings give a good estimate of the
 596 final in-use performance of the model; however, errors caused through
 597 the user interaction with such an application cannot be estimated
 598 from this data. The findings also highlight the vulnerability of the
 599 proposed method to a cumulative effect of errors made at previous
 600 presentations.

601 3.4 The decision making process of clinicians

602 To compare the algorithm described above with the clinical decision
 603 derived from the available data, a DT was trained on the data with
 604 relapse detection including the last five visits to approximate a data-
 605 derived rule to make predictions. This algorithm performed worse
 606 than the model described in section 3.1 with a F1 score of 0.688
 607 and 0.685 for the new and previously seen patients, respectively, and
 608 a Brier score of 0.437 and 0.418 for the new and previously seen
 609 patients respectively. This revealed the significant trade-off of the

610 more interpretable model with more historical data available compared
611 to the RF based approach. The decision tree showed that higher levels
612 within the tree are dose-based and hormone levels are considered
613 lower in the tree. This corresponds to the feature importance of the
614 RF selected as the final model in this study. This does not represent
615 a hardset rule on how the decision was derived, but is an indication of
616 how on average a decision could be made. Naturally, this will depend
617 on the clinician as well as the presenting patient on how the decision
618 is made in practice. A similar rule-based decision could be developed
619 by asking an experienced clinician to describe their decision making
620 process; however, this is unlikely to be a trivial task, as clinicians
621 with increasing amount of experience often use intuition or system 1
622 thinking (40; 41). This means that in practice neither the decision-
623 making process of a clinician nor a more advanced machine learning
624 algorithm can be fully understood.

625 4 Discussion and conclusion

626 We have demonstrated a machine learning-based approach to make
627 automated patient-centered *Carbimazole* dose predictions for hyper-
628 thyroidism patients. The results show good predictive performance
629 with acceptable reduction in accuracy in a simulated longitudinal
630 use case. This is, to our knowledge, the first example of machine
631 learning-based *Carbimazole* dosing. The algorithm developed could
632 be deployed as an application programming interface to be used by
633 patients via a mobile phone app. In its current form, the developed
634 model could reduce patient contacts by 50%; thus, potentially have
635 a significant positive impact on the workload of endocrine clinicians.

636 Through the removal of contact delays, this algorithm will also make
637 the turnaround times between blood test and a new dose prescription
638 much shorter. To ensure patient confidence in the method, a hybrid
639 between face-to-face contact with clinicians and the intermittent use of
640 the app would be desirable to ensure the safe use of the digital health
641 app. As this work is a pilot study, it is not ready for deployment in
642 its current form but is a starting point to continue the development
643 of the presented approach.

644 The shortcomings of this study are mainly the limited number of
645 patients from a single cohort, therefore future work would greatly
646 benefit from more data collected from other thyroid clinics to establish
647 a larger and richer dataset. Additionally, the data used in this study
648 was manually extracted from patient letters, however, if digital records
649 from an endocrinology clinic were accessible via automated means,
650 this could significantly speed up data collection and the further devel-
651 opment of the algorithm. A model-specific drawback of the presented
652 method is the use of previously predicted carbimazole doses, making
653 the model susceptible to accumulating errors. This problem could
654 be avoided by developing an algorithm that does not use previous
655 doses as input features or if the prediction target is changed to a
656 physiological outcome that could be used to make a dose prediction
657 based on the expected response of the patient. Both of these potential
658 solutions will be investigated in future work as a continuation of this
659 study.

660 This study showed that accurate prediction of *Carbimazole* doses
661 based on patient-centered data is possible and that such a digital health
662 solution could cut the number of required patient presentations by

⁶⁶³ half. This would not only have a positive impact on clinician workload,
⁶⁶⁴ but also result in financial savings for the healthcare providers.

665 Conflict of Interest Statement

666 The authors declare that the research was conducted in the absence
667 of any commercial or financial relationships that could be construed
668 as a potential conflict of interest.

669 Author Contributions

670 MB, TRi, and DK contributed to conception and design of the study.
671 DB extracted the patient data. JS developed the prototype app
672 and gave advise on technical issues. TRe, developed the algorithm,
673 analysed the data and wrote the first draft of the manuscript. MB, DK,
674 TRi, and RB critically analysed the findings. All authors contributed
675 to manuscript revision, read, and approved the submitted version.

676 Funding

677 This work was supported by a Higher Education Innovation Funding
678 fund.

679 References

- 680 1. Salvatore D, Davies TF, Schlumberger MJ, Hay ID, Larsen PR.
681 Thyroid Physiology and Diagnostic Evaluation of Patients With
682 Thyroid Disorders. *Williams Textbook of Endocrinology*. 2016:333–368.
- 683 2. Taylor PN, Albrecht D, Scholz A, *et al*. Global epidemiology of
684 hyperthyroidism and hypothyroidism. *Nature Reviews Endocrinology*.
685 2018;14(5):301–316.

- 686 3. Tunbridge WMG, Evered DC, Hall R, *et al.* THE SPECTRUM OF
 687 THYROID DISEASE IN A COMMUNITY: THE WHICKHAM
 688 SURVEY. *Clinical Endocrinology.* 1977;7(6):481–493.
- 689 4. De Leo S, Lee SY, Braverman LE. Hyperthyroidism. 2016.
- 690 5. Leigh S, Daly R, Stevens S, *et al.* Web-based internet searches
 691 for digital health products in the United Kingdom before and
 692 during the COVID-19 pandemic: A time-series analysis using app
 693 libraries from the Organisation for the Review of Care and Health
 694 Applications (ORCHA). *BMJ Open.* 2021;11(10):1–7.
- 695 6. Pagliari C. Digital health and primary care: Past, pandemic and
 696 prospects. *Journal of Global Health.* 2021;11:1–9.
- 697 7. Goodman B, Flaxman S. European union regulations on algorithmic
 698 decision making and a "right to explanation". *AI Magazine.*
 699 2017;38(3):50–57.
- 700 8. Cosgun E, Limdi NA, Duarte CW. High-dimensional pharmaco-
 701 genetic prediction of a continuous trait using machine learning
 702 techniques with application to warfarin dose prediction in African
 703 Americans. *Bioinformatics.* 2011;27(10):1384–1389.
- 704 9. Ma Z, Wang P, Gao Z, Wang R, Khalighi K. Ensemble of machine
 705 learning algorithms using the stacked generalization approach to
 706 estimate the warfarin dose. *PLoS ONE.* 2018;13(10).
- 707 10. Liu R, Li X, Zhang W, Zhou HH. Comparison of nine statistical
 708 model based warfarin pharmacogenetic dosing algorithms using the
 709 racially diverse international warfarin pharmacogenetic consortium
 710 cohort database. *PLoS ONE.* 2015;10(8):1–11.

- 711 11. Roche-Lima A, Roman-Santiago A, Feliu-Maldonado R, *et al.*
712 Machine learning algorithm for predicting warfarin dose in
713 caribbean hispanics using pharmacogenetic data. *The Lancet Oncol-*
714 *ogy*. 2019;244(1):395–406.
- 715 12. Roche-Lima A, Roman-Santiago A, Feliu-Maldonado R, *et al.* Ma-
716 chine learning algorithm for predicting warfarin dose in caribbean
717 hispanics using pharmacogenetic data. *Frontiers in Pharmacology*.
718 2020;10(January):1–8.
- 719 13. Asimwe IG, Blockman M, Cohen K, *et al.* Stable warfarin dose
720 prediction in sub-Saharan African patients: A machine-learning
721 approach and external validation of a clinical dose–initiation algo-
722 rithm. *CPT: Pharmacometrics & Systems Pharmacology*. 2022;11(1):20–
723 29.
- 724 14. Grossi E, Podda GM, Pugliano M, *et al.* Prediction of optimal war-
725 farin maintenance dose using advanced artificial neural networks.
726 *Pharmacogenomics*. 2014;15(1):29–37.
- 727 15. Jahmunah V, Chen S, Oh SL, Acharya UR, Chowbay B. Auto-
728 mated warfarin dose prediction for Asian, American, and Cau-
729 casian populations using a deep neural network. *Computers in Biology*
730 *and Medicine*. 2023;153:106548.
- 731 16. Tao Y, Chen YJ, Fu X, Jiang B, Zhang Y. Evolutionary en-
732 semble learning algorithm to modeling of warfarin dose predic-
733 tion for Chinese. *IEEE Journal of Biomedical and Health Informatics*.
734 2019;23(1):395–406.
- 735 17. Li Q, Tao H, Wang J, *et al.* Warfarin maintenance dose Prediction
736 for Patients undergoing heart valve replacement - A hybrid model

- 737 with genetic algorithm and Back-Propagation neural network.
738 2018.
- 739 18. Zaborek NA, Cheng A, Imbus JR, *et al.* The optimal dosing
740 scheme for levothyroxine after thyroidectomy: A comprehensive
741 comparison and evaluation. *Surgery (United States)*. 2019;165(1):92–
742 98.
- 743 19. Barrio M, Raeburn CD, McIntyre R, Albuja-Cruz M, Haugen BR,
744 Pozdeyev N. Computer-Assisted Levothyroxine Dose Selection
745 for the Treatment of Postoperative Hypothyroidism. *Thyroid®*.
746 2023;33(5):547–555.
- 747 20. Chen SS, Zaborek NA, Doubleday AR, *et al.* Optimizing Levothy-
748 roxine Dose Adjustment After Thyroidectomy With a Decision
749 Tree. *Journal of Surgical Research*. 2019;244:102–106.
- 750 21. Choi B, Jang JH, Son M, *et al.* Electrocardiographic biomarker
751 based on machine learning for detecting overt hyperthyroidism.
752 *European Heart Journal - Digital Health*. 2022;3(2):255–264.
- 753 22. Challa M. Optimized machine learning models for diagnosis and
754 prediction of hypothyroidism and hyperthyroidism. *Advances in*
755 *Intelligent Systems and Computing*. 2020;1108 AISC:286–293.
- 756 23. Duggal P, Shukla S. Prediction of thyroid disorders using advanced
757 machine learning techniques. *Proceedings of the Confluence 2020 -*
758 *10th International Conference on Cloud Computing, Data Science and*
759 *Engineering*. 2020:670–675.
- 760 24. Reddy K, Rani DJ. Comparison and Prediction of Hyperthy-
761 roidism Accuracy Rate Using Novel Deep Learning Technol-

- ogy and Vivo Monitoring. *Alinteri Journal of Agriculture Sciences*.
2021;36(1):698–703.
25. Shahajalal M, Rahman MM, Pranto SA, Ema RR, Islam T, Raihan M. Utilization of Machine Learning Algorithms for Thyroid Disease Prediction. 2022:57–69.
26. Aversano L, Bernardi ML, Cimitile M, *et al.* Thyroid Disease Treatment prediction with machine learning approaches. *Procedia Computer Science*. 2021;192:1031–1040.
27. Razia S, Siva Kumar P, Rao AS. Machine learning techniques for thyroid disease diagnosis: A systematic review. *Studies in Computational Intelligence*. 2020;885:203–212.
28. Azizi F, Abdi H, Mehran L, Amouzegar A. Appropriate duration of antithyroid drug treatment as a predictor for relapse of Graves’ disease: a systematic scoping review. *Journal of Endocrinological Investigation*. 2022;45(6):1139–1150.
29. Park S, Song E, Oh HS, *et al.* When should antithyroid drug therapy to reduce the relapse rate of hyperthyroidism in Graves’ disease be discontinued?. *Endocrine*. 2019;65(2):348–356.
30. Dutta P, Bhansali A, Walia R, Khandelwal N, Das S, Masoodi SR. Weight homeostasis & its modulators in hyperthyroidism before & after treatment with carbimazole. *The Indian Journal of Medical Research*. 2012;136(2):242.
31. Abbara A, Clarke SA, Brewster R, *et al.* Pharmacodynamic Response to Anti-thyroid Drugs in Graves’ Hyperthyroidism. *Frontiers in Endocrinology*. 2020;11(May):1–11.

- 787 **32. Davies TF, Laurberg P, Bahn RS. Hyperthyroid Disorders.**
788 *Williams Textbook of Endocrinology.* **2016:369–415.**
- 789 **33. Liu X, Qiang W, Liu X, *et al.* A second course of antithyroid drug**
790 **therapy for recurrent Graves’ disease: An experience in endocrine**
791 **practice. *European Journal of Endocrinology.* 2015;172(3):321–326.**
- 792 **34. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic**
793 **sampling approach for imbalanced learning. *Proceedings of the Inter-***
794 ***national Joint Conference on Neural Networks.* 2008(3):1322–1328.**
- 795 **35. Reza M, Miri S, Javidan R. RANDOM FORESTS Leo. *Machine***
796 ***Learning.*,. 2001;5(1):5–32.**
- 797 **36. Hand DJ. A Simple Generalisation of the Area Under the ROC**
798 **Curve for Multiple Class Classification Problems. tech. rep. 2001.**
- 799 **37. Niculescu-Mizil A, Caruana R. Predicting good probabilities with**
800 **supervised learning. in *Proceedings of the 22nd international conference***
801 ***on Machine learning - ICML ’05*(New York, New York, USA):625–**
802 **632ACM Press 2005.**
- 803 **38. Northcutt CG, Jiang L, Chuang IL. Confident learning: Estimating**
804 **uncertainty in dataset labels. *Journal of Artificial Intelligence Research.***
805 **2021;70:1373–1411.**
- 806 **39. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model**
807 **Predictions. in *Advances in Neural Information Processing Systems***
808 **(Guyon I, Luxburg UV, Bengio S, *et al.* , eds.);30Curran As-**
809 **sociates, Inc. 2017.**
- 810 **40. Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics**
811 **and Biases. *Science.* 1974;185(4157):1124–1131.**

- 812 **41. Roland D. Have we forgotten to teach how to think?. *Emergency***
813 *Medicine Journal.* **2017;34(2):68–69.**

	RMSE		F1	
Test set	Conventional	Compound	Conventional	Compound
new patients	3.862	4.435	0.751	0.681
previously seen patients	3.7	4.033	0.741	0.657

Table 1: Effect of longitudinal accumulation of errors

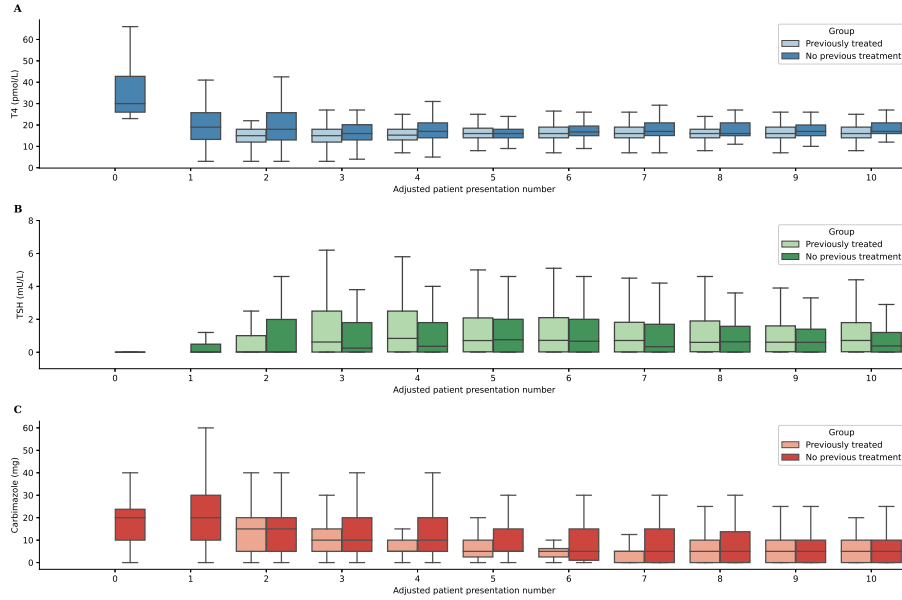


Figure 1: Comparison of the two patient groups following the date adjustment depending on whether a patient was previously treated somewhere else (outliers are not shown). (A) shows the T4 response by number of patient presentation. (B) shows the TSH response by number of patient presentation. (C) shows the *Carbimazole* dose by number of patient presentation.

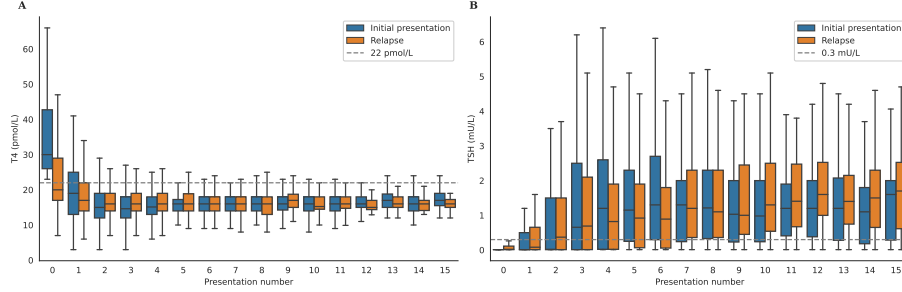


Figure 2: Comparison of the physiological response of patients at their initial presentation and patients when a relapse was observed. **(A)** shows the T4 response by number of patient presentation. **(B)** shows the TSH response by number of patient presentation.

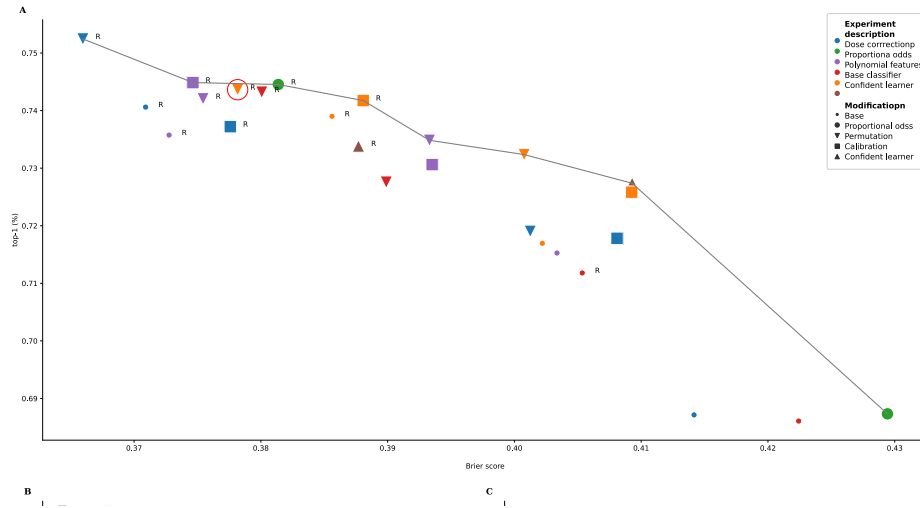


Figure 3: Comparison of the model performances using the Brier score to assess the model confidence and the top-1 to measure accuracy. Colours describe data regimes and marker shapes show the modification applied to the model. The grey line shows a Pareto front minimising the Brier score and maximising the top-1 score to highlight any possible candidate models. The model chosen for final deployment is highlighted with the red circle. **(A)** shows the the average values of the previously seen and new patient test set. **(B)** shows the the average values of the previously seen test set. **(C)** shows the the average values of the new patient test set.

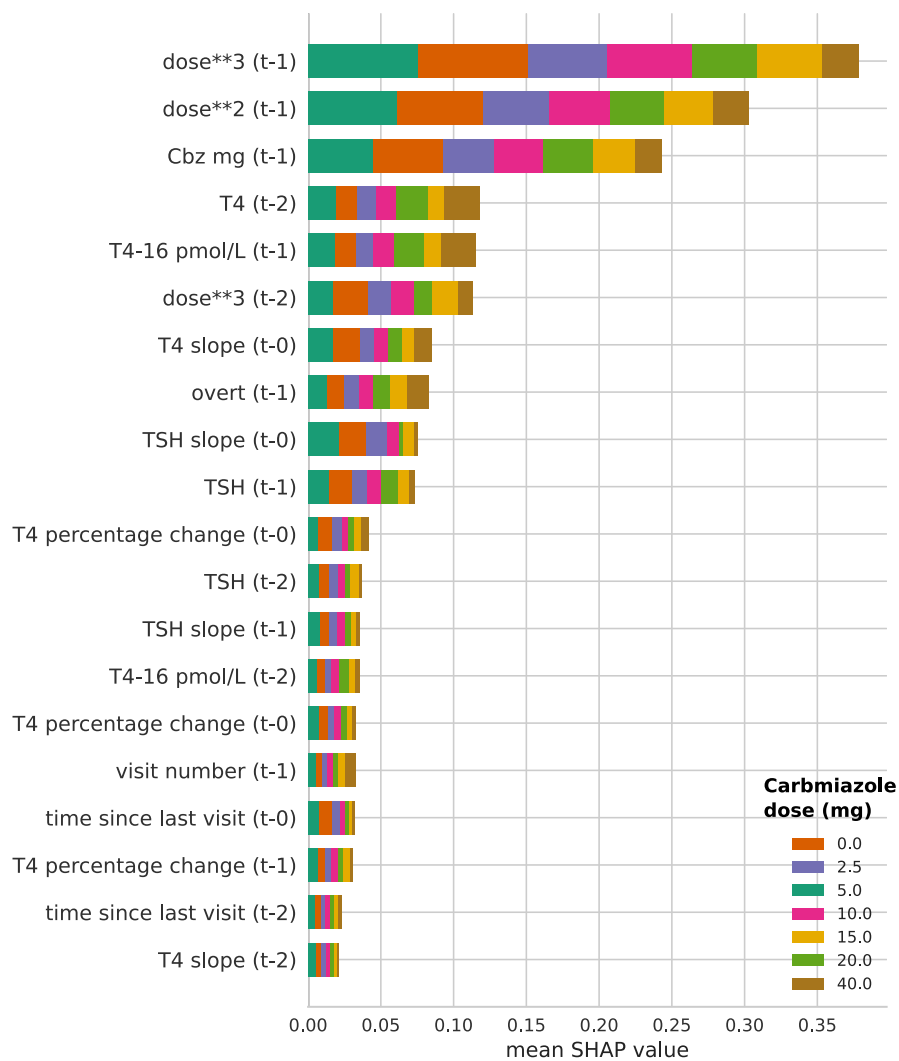


Figure 4: The average SHAP value for each feature and class. For features that are time dependent the time point is indicated in brackets, t, t-1 and t-2 indicate the value from the most recent, previous and the one before the previous presentations respectively compared to the target.

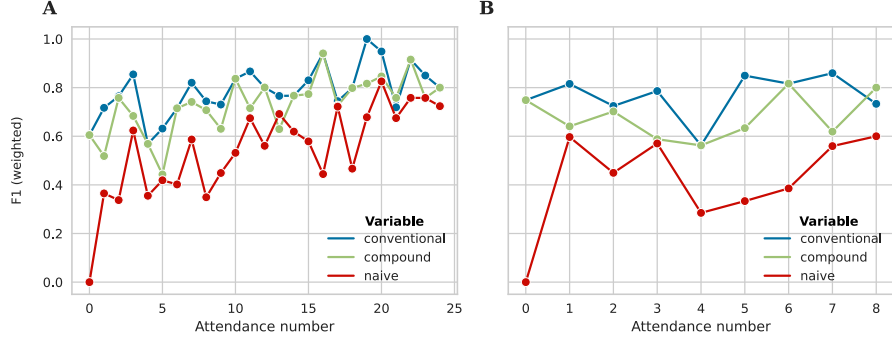


Figure 5: The F1 score for each of the sequential presentation numbers. **(A)** shows patients solely included in the test set, and **(B)** shows patients whose first half of presentations were included in the training set.

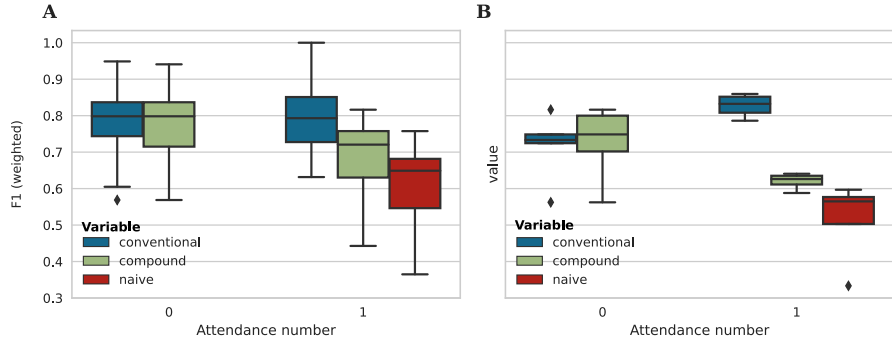


Figure 6: The absolute errors in mg score for each of the sequential presentation numbers and predictions. **(A)** shows patients solely included in the test set, **(B)** shows patients whose first half of presentations were included in the training set.

Algorithm	Parameter	Values
Logistic regression	penalty	none, L2
KNN	n	1-20
SGD	alpha	0.0001, 0.0005, 0.001, 0.005, 0.01
Extra Trees	n	10-200
	criterion	gini, entropy
RF	n	10-200
	criterion	gini, entropy
DT	criterion	gini, entropy
	splitter	random, best

Table 2: Algorithms used and gridsearch parameters

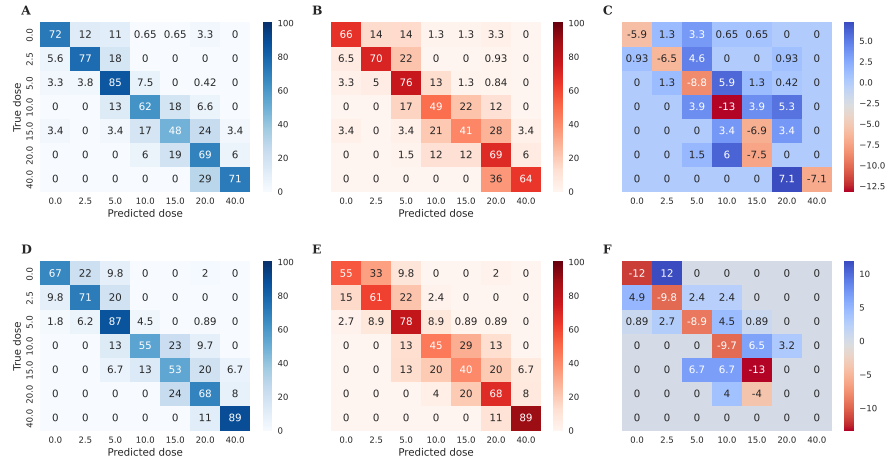


Figure 7: The confusion plots and the residual to compare between the conventional and compound prediction and their residuals. **A-C** show patients solely included in the test set. **(A)** shows the conventional error, **(B)** shows the compound error, **(C)** residuals (compound - conventional). **D-F** show patients whose first half of presentations were included in the training set. **(D)** shows the conventional error, **(E)** shows the compound error, **(F)** residuals (compound - conventional)

			New				Seen				
	Variety	Relapse	F1	Brier	ROC	top-1	F1	Brier	ROC	top-1	
Base dataset	base		0.68	0.454	0.929	0.677	0.761	0.35	0.96	0.757	
		X	0.73	0.396	0.944	0.724	0.754	0.375	0.955	0.754	
	permutation		0.693	0.45	0.928	0.692	0.778	0.351	0.958	0.773	
		X	0.751	0.38	0.945	0.748	0.74	0.377	0.951	0.739	
	calibration		0.691	0.471	0.929	0.688	0.766	0.348	0.958	0.763	
		X	0.744	0.383	0.944	0.741	0.744	0.393	0.95	0.743	
Gap-removal dataset	base		0.677	0.457	0.927	0.673	0.766	0.35	0.959	0.758	
		X	0.719	0.395	0.944	0.714	0.757	0.351	0.96	0.757	
	permutation		0.685	0.452	0.923	0.681	0.792	0.335	0.956	0.788	
		X	0.739	0.392	0.944	0.731	0.753	0.359	0.957	0.753	
	calibration		0.676	0.462	0.926	0.673	0.79	0.325	0.955	0.788	
		X	0.736	0.388	0.943	0.733	0.758	0.361	0.955	0.757	
Dose-correction dataset	base		0.646	0.465	0.925	0.64	0.739	0.363	0.955	0.735	
		X	0.727	0.397	0.943	0.719	0.76	0.345	0.962	0.762	
	permutation		0.691	0.446	0.929	0.684	0.754	0.357	0.954	0.754	
		X	0.743	0.382	0.945	0.738	0.766	0.35	0.956	0.767	
	calibration		0.677	0.463	0.927	0.674	0.763	0.353	0.949	0.762	
		X	0.729	0.396	0.94	0.726	0.749	0.36	0.954	0.749	
Polynomial features	base		0.657	0.464	0.924	0.652	0.723	0.381	0.951	0.72	
		X	0.703	0.424	0.933	0.698	0.727	0.387	0.949	0.725	
	permutation		0.693	0.447	0.928	0.682	0.776	0.333	0.959	0.773	
		X	0.747	0.384	0.946	0.743	0.743	0.376	0.953	0.743	
Confident learner				0.696	0.461	0.925	0.695	0.77	0.357	0.955	0.76
			X	0.739	0.383	0.941	0.732	0.741	0.392	0.943	0.736
Ordinal predictor				0.69	0.45	0.919	0.684	0.699	0.409	0.931	0.691
			X	0.748	0.389	0.936	0.739	0.752	0.374	0.947	0.75

Table 3: Summary of experimental results variety indicates the data variation or the special predictor, relapse indicates the application of relapse-detection to the datasets marked as "X".