Supplementary data

1 Supplementary methods

The following sections describe the feature engineering of the data in more detail aimed at the technical reader.

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

1.1 Binning of doses

As clinicians sometimes prescribed doses on alternate days or only a few times per week, several doses were rarely, or in some cases only prescribed once. Such doses were impossible to predict as a classification problem, and, therefore a binning method was used, which assigned such doses to the closest user-defined dose. If a dose fell exactly in the middle of two neighbouring doses, the lower dose was selected. This is because patients tend to prefer lower doses, as these will result in less weight gain (1). The doses used for this binning process were based on the actual prescription pattern and were chosen in consultation with a clinician. The selected binned doses were: 0, 2.5, 5, 10, 15, 20, 40 mg.

1.2 Dose exponents

A feature suggested in the literature was dose exponents, which are calculated by squaring and cubing the previous doses. These have previously been shown to improve the accuracy of the prediction (2).

1.3 Time between presentations

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

1.4 Thyroid hormone based features

It was suggested by (2) that an important feature is the daily change in T4 level. This was calculated based on the time between presentations and the change in recorded T4 levels. Furthermore, the percentage change of T4 was calculated to give an indication of how fast the change occured. Additionally, the difference to the normal range was explicitly added, which meant that the approximate mean target (16 pmol/L) was subtracted from the recorded reading. In addition, the percentage change in TSH was calculated to give an indication of how quickly the change occured.

1.5 Overt andsubclinical thyroidism

Clinically, hyperthyroidism is distinguished into overt and subclinical states (3). To ensure that this information was available to the model, Booleans were added for both conditions. For this purpose, we used the following definitions:

- Overt: TSH <0.5 mU/L and T4 >21 pmol/L
- Subclinical: TSH <0.5 mU/L and 10 <>T4 <21 pmol/L

1.6 T3

The levels of T3 are not routinely tested in current practice. This meant that these were not included for each presentation. However, T3 was often recorded once at the beginning of a patient's treatment. Therefore, the ratio of T4:T3 was included as a static patient characteristic, as this can provide additional information on the underlying pathophysiology of hyperthyroidism (2). Missing values were imputed to the median.

1.7 Imputation

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

1.8 Manual time adjustments

Due to the complexity of the dataset, which included timeseries data for each patient, identifying spurious data was challenging. Therefore, a manual screen was performed to find breaks in the timeline of individual patients. Several of these cases were identified;

most were patient contacts in unusually rapid succession. As these mostly followed a thyroid function test with a long follow-up period, it was concluded in consultation with a clinician that this was probably due to a repeat of the test at the next presentation because the follow-up was not timely. In these cases, the initial readings with delayed follow-up were removed because the clinician would have made their decision based on the more recent repeated blood test.

1.9 Patient grouping

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

It was assumed that the time in treatment was an important factor for making an accurate prediction. Therefore, it was expected that the time in treatment for patients who previously received treatment was a confounding factor. To adjust the data to reflect previous treatments for these patients, starting treatment was adjusted to begin at the third presentation where a patient was expected to have T4 levels within the normal range. Furthermore, the average time to the third presentation was used as a starting point. These modifications to the data based on domain knowledge made the prediction task easier as patients at physiologically different stages of treatment were distinguished; see Figure S1.

1.10 Relapse detection

The aim of antithyroid treatment is to normalise thyroid function with the lowest dose of Carbimazole with a view to trial a cessation of treatment. However, many patients suffer a relapse which requires an adjustment of their Carbimazole dose at some point during their treatment time. In this study, we defined a relapse as a TSH level within the normal range (>0.3 mU/L) followed by one below the normal range of 0.3 mU/L. This applied to

61.7% of our patients, which was within the range of the current literature (4). The relapse episodes were labeled with a Boolean. This was important as the treatment of a relapse is different from the initial treatment (Figure 2). Part of the reason for this is that relapses were identified faster due to routine tests compared to the initial diagnosis.

1.11 Polynomial features

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

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

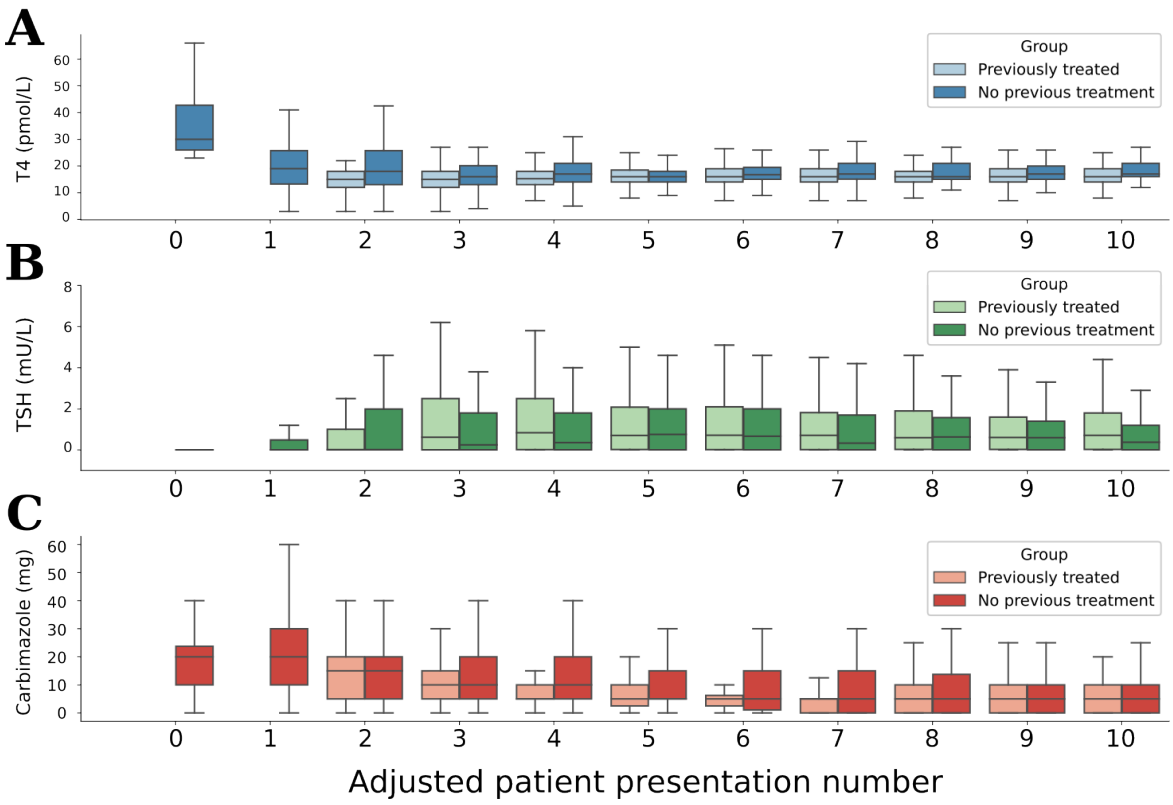

Figure S1: Comparison of the two patient groups following the date adjustment depending on whether a patient was previously treated somewhere else (outliers are not shown). (A) shows the T4 response by number of patient presentation. (B) shows the TSH response by number of patient presentation. (C) shows the Carbimazole dose by number of

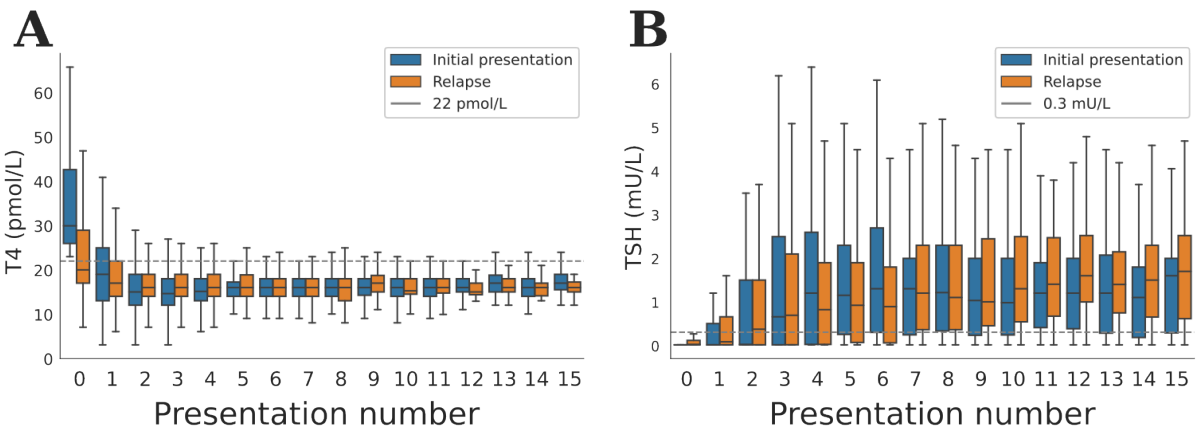

Figure S2: Comparison of the physiological response of patients at their initial presentation and patients when a relapse was observed. (A) shows the T4 response by number of patient presentation. (B) shows the TSH response by number of patient presentation.

2 Data descriptives

As hyperthyroidsm is more common in female patients, a skewed data set was expected. Our dataset contained 77.8% female patients. The age distributions were similar for both sexes with a mean age at first admission of 52.4 and 54.1 years for the female and male patients, respectively (Figure S1 and Table S1).

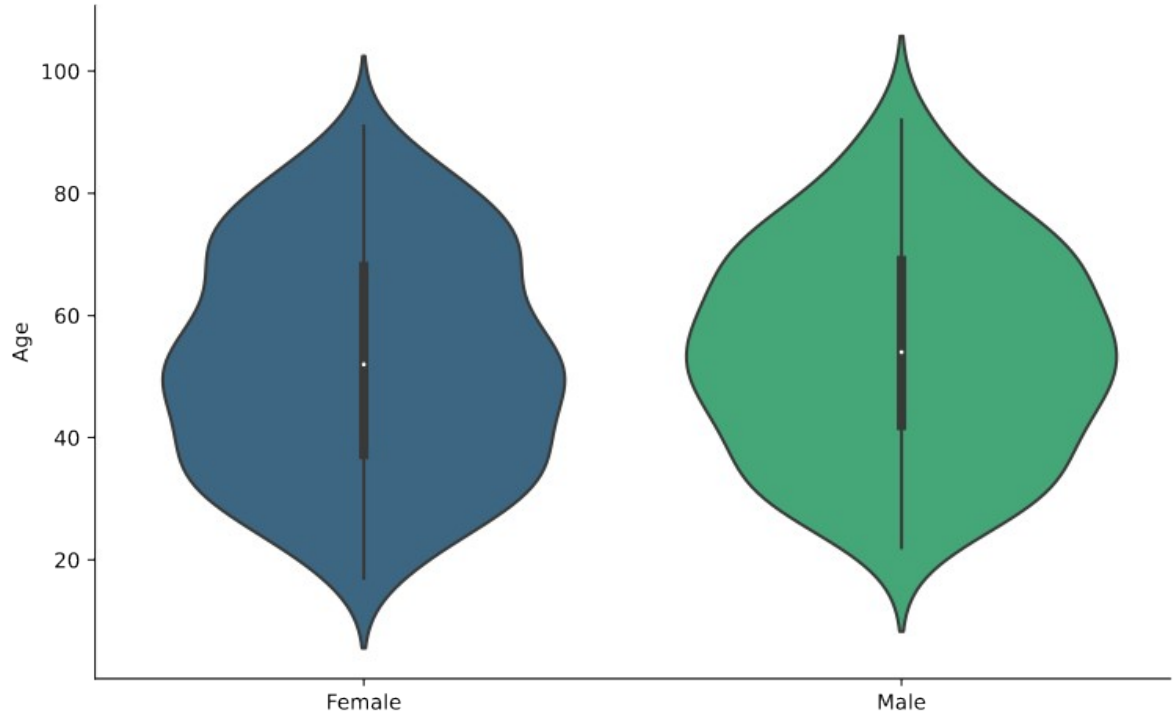

Figure S3: Age distributions at the first presentation for both sexes. Note that the male patients are underrepresented at 22.2% counts: female=326, male=93)

The average treatment time was 4 years with a maximum of 20 years, highlighting the wide range of treatment times. Some patients did not attend for very long periods within the total treatment period. The majority (61.7%) of the patients suffered one or more relapses during their treatment.

	Female	Male
n (%)	326 (77.8)	93 (22.2)
Age (mean (std[1]))	52.4 (18.2)	54.1 (17.0)
Years in treatment (mean (std))	4.1 (3.7)	3.7 (3.3)
Presentation counts (mean (std))	15.9 (11.5)	13.2 (10.4)
>=1 relapses (n (% of total))	186 (44.1)	48 (11.4)
Time in days between presentations (mean (std))	94.7 (186.6)	102.8 (232.7)
Overt (n total presentations (%))	1022 (15.9)	262 (4.1)
Sub-clinical (n total presentations (%))	1440 (22.4)	235 (3.7)
Exclusions (n (%)):		
<4 presentations	7 (1.7)	5 (1.2)
Pregnancy	14 (3.3)	-
Free text	16 (3.8	6 (1.4)
Long term high dose	5 (1.2)	0
Dose >40mg	10 (2.4)	3 (0.7)
Included (n (% of total))	279 (66.1)	74 (17.5)

Table S1: Summary statistics for raw data.

Supplementary References:

1. 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.
2. Abbara A, Clarke SA, Brewster R, et al. Pharmacodynamic Response to Anti-thyroid Drugs in Graves' Hyperthyroidism. Fron-tiers in Endocrinology. 2020;11(May):1–11.
3. National Health Service England, & National Health Service Improvement. (2022, May). *Implementing patient initiated follow-up*. https://www.england.nhs.uk/wp-content/uploads/2022/05/B0801-implementing-patient-initiated-follow-up-guidance-1.pdf
4. Liu X, Qiang W, Liu X, et al. A second course of antithyroid drug therapy for recurrent Graves' disease: An experience in endocrine practice. European Journal of Endocrinology. 2015;172(3):321–326.

[1] Standard deviation