Original Article

JHSMR Journal of Health Science and Medical Research

Long-Term Evaluation of Allopurinol Effectiveness in the Thai Population: Insights from 10 Years of Real-World Big Data Analysis

Kanoot Jaruthamsophon, M.D., Ph.D.[1], Polathep Vichitkunakorn, M.D., Ph.D.[2], Warit Ruanglertboon, B.Pharm., M.Sc., Ph.D.[3]

[1]Department of Pathology, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand.
[2]Department of Family and Preventive Medicine, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand.
[3]Discipline of Pharmacology, Division of Health and Applied Sciences, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand.

Received 28 February 2024 ● Revised 22 April 2024 ● Accepted 23 April 2024 ● Published online 19 September 2024

Abstract:

Objective: This study aimed to evaluate the effectiveness and safety profile of allopurinol in Songklanagarind hospital, Thailand.

Material and Methods: Electronic medical records (EMRs) of patients having undergone allopurinol treatment for gouty arthritis, or other pertinent indications, were retrospectively reviewed. The primary outcome of this study was the longitudinal measurement of uric acid levels. A linear mixed model (LMM) analysis assessed the statistical significance of uric acid reduction across hospital visits. The safety profile regarding allopurinol-related adverse drug reactions was assessed.

Results: A total of 11,277 EMRs were extracted, of which 8,801 EMRs were eligible for further analyses. These records pertain to a cohort of 3,219 unique patients whose visits were arranged chronologically from the earliest to the fifth recorded visit. Overall, uric acid level decreased after allopurinol treatment in all patients. Across subsequent visits, the analysis, using LMM, found a significant reduction in uric acid levels (χ^2=1465.4, df=4, p-value<0.001), with the early reduction observed as early as the second visit. The incidence of allopurinol-related adverse reactions was 4.37% (385/8801 cases).

Conclusion: The current dosing regimen of allopurinol at Songklanagarind Hospital has demonstrated favorable treatment outcomes and continues to be effective. This data provides insights into pharmacological management and its real-world outcomes.

Keywords: allopurinol, clinical pharmacology, real-world data, treatment outcome, uric acid

Contact: Warit Ruanglertboon, B.Pharm., M.Sc., Ph.D.
Discipline of Pharmacology, Division of Health and Applied Sciences, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand.
E-mail: warit.r@psu.ac.th

J Health Sci Med Res 2025;43(2):e20241087
doi: 10.31584/jhsmr.20241087
www.jhsmr.org

Introduction

Allopurinol is a well-established medication commonly used to treat several conditions associated with elevated levels of uric acid; including hyperuricemia, gouty arthritis, and kidney stones[1]. Importantly, its efficacy has been proven through extensive global practice in effectively controlling the levels of uric acid[2]. Therefore, allopurinol is regarded as the primary urate-lowering medication endorsed by the US Food and Drug Administration (FDA), and numerous medical associations[2,3].

Ranked as the fourth most densely populated country in Southeast Asia, Thailand has faced a variety of challenges, among which are significant disparities in household income. Consequently, these challenges also influence healthcare expenditure, resulting in diverse levels of access and affordability for medical services among individuals with varying income levels. In relation to this issue, allopurinol has found strong acceptance within the Thai healthcare community due to its accessible cost and proven effectiveness. Febuxostat, another xanthine oxidase inhibitor, stands as an alternative first-line option to allopurinol[4,5], which is now available in Thailand. Nevertheless, when compared to allopurinol, there is a substantial difference in cost; with febuxostat being notably more expensive. Therefore, while the efficacy of these two medicines is comparable, taking the cost factor into account, allopurinol remains a more practical choice for Thai society other than febuxostat.

Although the well-documented benefits of allopurinol across multiple dimensions are recognized, the occurrence of concomitant adverse events during its utilization has presented a challenge. The Asian population has exhibited a notable incidence of severe skin and cutaneous reactions (SCAR); for example: Stevens-Johnson syndrome (SJS), drug reactions with eosinophilia and systemic symptoms (DRESS) as well as other events linked to exposure to allopurinol. The mechanism strongly tied to these events is believed to be associated with distinct genetic polymorphisms of the *HLA-B* gene; specifically the *HLA-B*58:01* variant[6,7]. Research conducted in Thailand has demonstrated that individuals that have encountered the emergence of SCARs after allopurinol administration have a notably higher frequency of the *HLA-B*58:01* genetic variant.

The associated odds ratio was calculated at 348.3, with a corresponding 95% confidence interval risk; spanning from 19.2 to 6336.9[8]. Furthermore, numerous studies have also observed a similar pattern of observation, implying the need for increased awareness and caution in the prescription of allopurinol within medical practices[9,10]. The recognition of risks associated with allopurinol, particularly for patients about to begin treatment, has led to the adoption of a strategy known as: "start low, go slow." This involves administering an initial dosage of 100 mg/day allopurinol, rather than 300 mg/day allopurinol in patients with normal kidney function along with slow escalation of allopurinol dosage every 2–4 weeks, if required[11]. However, to ascertain whether this reduced dose is sufficiently effective, it is essential first to establish a basic understanding of its overall efficacy. Recently, there has been a noticeable lack of systematic reviews assessing the efficacy of allopurinol among the Thai population; especially in the southern regions where the ethnic composition varies significantly from other areas[12,13]. Crucially, a comprehensive understanding of treatment efficacy, patient characteristics, and related paradigms is vital for the advancement of pharmacotherapy. This knowledge is key to fully implementing improved prescribing patterns for allopurinol.

Recent technological advancements have enabled various disciplines to utilize updates and gain a competitive edge in knowledge development, one of which is the field of data science[14,15]. Interestingly, the medical research field has also adopted this trend, resulting in the emergence of a wide range of profound understanding within certain

disciplines[16,17]. One of the current focal points in the research methodology of data science is the analysis of big data. In reality, big data is frequently encountered on a daily basis, particularly within hospital settings. Patient record data has transitioned to digital health records, with information typically stored within hospital databases. As such, utilizing the analytics approach to big data at the hospital level would facilitate the derivation of invaluable insights and contribute to evidence-based decision-making as well as enhancements in patient care[18,19].

Therefore, this current study aimed to retrospectively analyze hospital-level data at Songklanagarind Hospital, focusing primarily on assessing the effectiveness of allopurinol across various paradigms; including the corresponding safety data. This study emphasizes the use of statistical approaches as the foundational methodology, complemented by additional analytical techniques. The outcome of the study is expected to showcase the utility of these comprehensive analytical methods in addressing the intricate and unstructured nature of real-world data, disclosing hidden insights within the extensive datasets.

Material and Methods

Study design and setting

This current study was a retrospective cohort study conducted at Songklanagarind Hospital, Hat Yai, Songkhla, Thailand. Electronic medical records (EMRs) of patients having received allopurinol for the treatment of gouty arthritis or other relevant indications were extracted. The data was collected from 2010 to 2022, with technical assistance provided by the Division of Digital Innovation and Data Analytics (DIDA), Faculty of Medicine, Prince of Songkla University. This research has received ethical approval from the Office of Human Research Ethics Unit, Faculty of Medicine, Prince of Songkla University to ensure compliance with ethical guidelines and standards (REC. 65-208-19-2).

Study variables and outcomes

The collected variables from the EMR included: gender, age, uric acid levels, dose of prescribed allopurinol, allergic history, and laboratory data. The main inclusion criterion for data retrieval was patients having received allopurinol at Songklanagarind hospital from 2010 to 2022, with recorded uric acid level data, and whom had more than one hospital visit.

The primary outcome of this study was the longitudinal measurement of uric acid levels. Therefore, patients with no subsequent hospital follow-up were subsequently excluded from the study. As part of a subgroup analysis, the study also presents a summarized account of the safety profile concerning reported adverse events associated with allopurinol. Additionally, the diagnosis of allopurinol-related ADRs was meticulously extracted. Only those ADRs evaluated by pharmacists, or with the collaborative assistance of pharmacists were included for detailed analysis and enumeration in the study. Furthermore, supplementary analysis was conducted to examine the distribution of the prescribed allopurinol dose within the cohort.

Data processing

The EMR data obtained from DIDA was initially observed to lack a coherent and organized structure. To address the issue, the 'data.table' and 'tidyverse' packages in R software were used for formatting all variables. This resulted in a well-structured data frame that included baseline characteristics, uric acid levels, allopurinol doses, and safety profiles. After the initial formatting process, the data concerning baseline characteristics and laboratory measurements, with a particular focus on uric acid levels, were merged. Consequently, this merging resulted in the identification of several patients whose uric acid levels were found to be missing, leading to their subsequent exclusion from the dataset. The total number of recruited EMR data was 11,277 records.

In the process of preparing data for major analysis, the authors undertook several steps; including data transformation and feature selection, to ensure the data's integrity and relevance. Age grouping and gender were chosen as variables based on existing research that highlights their significant influence on uric acid levels. Age groups were categorized into: 'Baby and adolescents (aged 0–17)', 'Young adults (18–34)', 'Middle-aged adults (35–50)', and 'Older adults (>50)', to accurately reflect the overall patient profiles in this study. This categorization considers the variation in uric acid metabolism and excretion as age progresses.

Following the initial step, data transformation was carried out. An example of this can be seen in the categorization of uric acid levels, which were divided based on a clinically significant cut-off point of 7.2 mg/dL: adhering to the guidelines of Songklanagarind Hospital. This threshold was used to classify patients into two groups: those with uric acid levels below 7.2 mg/dL and those with levels equal to or above 7.2 mg/dL. Similarly, the labelling of responses regarding the reduction in uric acid levels due to treatment was conducted during this preparation stage. This was to facilitate subsequent classification analysis using specific statistical approaches.

Hospital follow-up of patients was labeled according to the order of time each patient visited the hospital. This was referred to as: 'visit,' with five subsequent visits marked (1st visit to 5th visit), so as to serve as time points for longitudinal analysis. This temporal categorization aids in understanding how treatment approaches and patient outcomes have evolved throughout the study period. The descriptive analysis of the data, following data preprocessing, is presented in Table 1. For detailed technical information on data processing, please refer to the: 'Technical declaration regarding data processing', within the Supplementary section.

Statistical analysis

Descriptive statistics were utilized to demonstrate baseline characteristics, the dose of prescribed allopurinol, median with interquartile range (IQR) follow-up time between patients' visits, uric acid levels, and safety profile. Additional descriptive analysis was also summarized as mean, standard deviation (S.D.), standard error (SE) and percentage. As the dataset in this analysis is reasonably large, the Central Limit Theorem has been applied to justify the use of parametric methods for specific analyses. However, descriptive statistics are presented using both means with standard deviation (S.D.) and medians, as exemplified by the uric acid levels and follow-up times, to ensure a clear and accurate representation of the data's central tendency and variability. This dual approach addresses potential concerns in regard to the differences between these statistical descriptors.

A linear mixed model (LMM) was used to describe the difference in uric acid levels across follow-up visits, wherein subjects were modeled as a random effect and hospital visits were modeled as a fix effect. Post-hoc analysis was performed using Tukey's HSD to identify the statistical differences across visits. Data processing up to the analysis was conducted using R version 4.1.1, with the necessary computing packages.

Building on the initial statistical analyses, a supervised machine learning process was proposed. This process aims to assess the feasibility of predicting the outcomes of uric acid reduction via leveraging variables that have been further refined and are currently available after comprehensive data processing.

Results

Patients' characteristics across the working datasets

Due to the longitudinal nature of the analysis, the subsequent investigation focused on patients that had more than one visit; enabling the measurement of changes in

the level of uric acid over time. Hence, out of the 11,277 records obtained from the initial recruitment, only 8,801 records were available for the actual analysis.

The patients' visits were categorized into five time points, which were arranged chronologically from the earliest visit to the fifth recorded visit. The numbers of patients in each visit were 3,219, 3,219, 1,014, 334, and 1,015, respectively. Herein, the main focus of the analysis was on four consecutive visits, as the data of subsequent visits was available only in a small number of patients, thus limiting the ability to achieve reliable comparisons. Therefore, visits that occurred after the fourth consecutive visit were combined and labeled as the fifth visit and later for the purpose of this analysis.

Changes in uric acid levels

Table 1 presents the distribution of patients' characteristics across various visit ranges. The first two visits exhibited the highest patient density, while visits three and four demonstrated a gradual decrease in the number of patients. Male patients constituted the predominant gender classification in this dataset, and middle-aged adults (age 35–50 years old) and older adults (age >50 years old) accounted for more than 80% of the total patient population (Supplementary Figure S2). There were 7 cases with an age of less than 1 year old that were incidentally included for the treatment of tumor lysis syndrome. These cases were, therefore, excluded from further analyses. The distribution of patients based on years of enrollment (period of laboratory visits, Table 1) was evenly spread across the entire ten-year range; indicating a consistent epidemiology of the disease.

Initially, in this patient cohort, the uric acid level was 8.80 [7.50–10.20], while in the subsequent visits, there was a progressive reduction (Table 1). On the second visit, it was 7.40 [6.00–8.70], and in the third visit, it further decreased to 6.90 [5.40–8.60]. Correspondingly, as the uric acid levels gradually changed during the subsequent visits, there were also variations in the proportion of patients initially exceeding the cut-off point of 7.2; indicating hyperuricemia. Therein, a substantial portion of patients were classified as having hyperuricemia; comprising of 80.1% of the cohort. However, over the entire follow-up period within the dataset range, this proportion dramatically decreased to 36% at the fourth visit. Furthermore, if subsequent visits were included, the proportion decreased even further to 20%.

Median follow-up time was also been calculated and is summarized in Table 2. The study participants exhibited extended gaps between their initial visits, while follow-up durations progressively shortened in subsequent visits.

Graphical presentations for the overall trend of uric acid levels and subgroup analysis by age group are shown in Figures 1 and 2, respectively. During the course of treatment, a small number of patients exhibited an increase in the level of uric acid despite receiving the treatment. Supplementary Figure 3 displays the types of responses, quantified by the number of patients, following allopurinol treatment.

LMM analysis: reduction of uric acid across multiple visits

The analysis of deviance, as part of the LMM, revealed a highly significant effect of hospital follow-up on uric acid levels (χ^2=1465.4, df=4, p-value<0.001). This suggests that individuals whom continuously take medicine or, at the very least, return for follow-up tend to demonstrate substantially different uric acid levels. Table 3 depicts a fixed effects analysis from the LMM model, examining uric acid levels as the response variable across multiple visits. The intercept value of 8.7643 represents the estimated level of uric acid at the baseline, corresponding to Visit 1. Subsequent visits from the second visit onwards exhibit a consistent decrease in uric acid levels; as detailed by the negative values in the 'Value' column. Furthermore, the corresponding 't-values', which quantify the ratio of the

estimated effect to its standard error, elucidate the strength of the data in supporting an actual effect versus the null hypothesis of no effect. Higher absolute t-values signify more substantial evidence against the null hypothesis, implying that the observed changes are not merely due to random variation. This analysis, therefore, reveals a correlation between the negative values (indicating decrease in uric acid levels from the second visit onwards) and the negative t-values, affirming a reduction in uric acid levels.

In the random effects analysis presented in Table 4, the residual standard deviation within subjects, after adjusting for the fixed effects in the model, is reported to be 1.96. This value signifies the variability in uric acid levels within subjects across various visits, which remains unexplained by the fixed effects previously mentioned. Furthermore, Table 5 shows an analysis of covariance parameters, illustrating the relationships between subsequent visits concerning changes in uric acid levels. Notably, the third visit displays a correlation coefficient of 0.323 with the second visit, indicating a positive correlation at the lower end of the moderate range in terms of their impact on uric acid levels. This suggests that individuals that experienced a certain level of reduction in uric acid levels during the second visit are likely to show a similar trend of reduction during the third visit. However, this relationship, while indicating a consistent direction of change between these visits, is not strongly predictive.

Table 1 Patients' characteristics across five visits

Patient demographics and clinical characteristics	First visit (N=3,219)	Second visit (N=3,219)	Third visit (N=1,014)	Fourth visit (N=334)	Fifth visit and later (N=1,015)	Overall (N=8,801)
Gender						
Female	671 (20.8%)	671 (20.8%)	255 (25.1%)	107 (32.0%)	467 (46.0%)	2,171 (24.7%)
Male	2,548 (79.2%)	2,548 (79.2%)	759 (74.9%)	227 (68.0%)	548 (54.0%)	6,630 (75.3%)
Age group (years)						
Baby and adolescence (aged 0–17)	2 (0.1%)	2 (0.1%)	1 (0.1%)	1 (0.3%)	1 (0.1%)	7 (0.1%)
Young adults (18–34)	360 (11.2%)	356 (11.1%)	142 (14.0%)	70 (21.0%)	307 (30.2%)	1,235 (14.0%)
Middle-aged adults (35–50)	1,165 (36.2%)	1,155 (35.9%)	368 (36.3%)	122 (36.5%)	401 (39.5%)	3,211 (36.5%)
Older adults (>50)	1,692 (52.6%)	1,706 (53.0%)	503 (49.6%)	141 (42.2%)	306 (30.1%)	4,348 (49.4%)
Uric acid level (mg/dL)						
Median [IQR]	8.80 [7.50, 10.20]	7.40 [6.00, 8.70]	6.90 [5.40, 8.60]	6.20 [4.70, 8.20]	4.80 [3.40, 6.60]	7.70 [5.80, 9.30]
Mean (S.D.)	8.76 (2.56)	7.33 (2.39)	7.07 (2.49)	6.54 (2.81)	5.36 (2.99)	7.57 (2.77)
Range of uric acid level (mg/dL)						
<7.2	640 (19.9%)	1,471 (45.7%)	532 (52.5%)	205 (61.4%)	812 (80.0%)	3,660 (41.6%)
≥7.2	2,579 (80.1%)	1,748 (54.3%)	482 (47.5%)	129 (38.6%)	203 (20.0%)	5,141 (58.4%)
Period of laboratory visits						
<2010	875 (27.2%)	831 (25.8%)	229 (22.6%)	66 (19.8%)	80 (7.9%)	2,081 (23.6%)
2010–2013	586 (18.2%)	614 (19.1%)	215 (21.2%)	79 (23.7%)	273 (26.9%)	1,767 (20.1%)
2013–2016	533 (16.6%)	528 (16.4%)	180 (17.8%)	58 (17.4%)	149 (14.7%)	1,448 (16.5%)
2016–2019	504 (15.7%)	503 (15.6%)	165 (16.3%)	57 (17.1%)	178 (17.5%)	1,407 (16.0%)
2019–early 2022	721 (22.4%)	743 (23.1%)	225 (22.2%)	74 (22.2%)	335 (33.0%)	2,098 (23.8%)

IQR=interquartile range; S.D.=standard deviation; mg=milligram; dL=deciliter
Tests of normality using the Shapiro-Wilk and Kolmogorov-Smirnov tests indicate a non-normal distribution of uric acid levels across visits. Given the large size of the dataset, the Central Limit Theorem has been applied to justify the use of parametric methods. However, data are described using both means and medians to ensure a clear and accurate representation of the data's central tendency and variability, so as to address potential concerns regarding the differences between these descriptors.

Table 2 Follow-up time in days between visits (from the second visit onward).

Follow-up time between consecutive visits	Female (n=671)	Male (n=2,548)	Overall (n=3,219)
Follow-up time between the first and second visit (days)			
Median [IQR]	35.0 [21.0, 63.0]	42.0 [28.0, 68.0]	42.0 [28.0, 66.0]
Mean (S.D.)	42.5 (27.6)	47.7 (25.8)	46.6 (26.3)
Follow-up time between the first and second visit (days)	Female (n=255)	Male (n=759)	Overall (n=1,014)
Median [IQR]	28.0 [7.0, 36.5]	34.0 [21.0, 49.0]	29.5 [19.0, 48.0]
Mean (S.D.)	26.8 (20.0)	34.1 (19.1)	32.3 (19.5)
Follow-up time between the third and fourth visit (days)	Female (n=107)	Male (n=227)	Overall (n=334)
Median [IQR]	5.00 [1.0, 27.0]	21.0 [4.0, 30.0]	14.0 [2.0, 28.0]
Mean (S.D.)	12.3 (13.4)	20.1 (15.6)	17.6 (15.4)

IQR=interquartile range; S.D.=standard deviation

Table 3 LMM analysis on the fixed effects for the level of uric acid across different visits

	Fixed effect	Value	Std. error	DF	t-value	p-value
	Intercept	8.7643	0.0443	5578	197.8384	<0.0001
	Second visit	−1.4295	0.0489	5578	−29.2422	<0.0001
	Third visit	−1.7438	0.0756	5578	−23.0596	<0.0001
	Fourth visit	−1.9879	0.1230	5578	−16.1577	<0.0001
	Fifth visit and later	−3.0092	0.1078	5578	−27.9217	<0.0001

DF=degrees of freedom, LMM=linear mixed model

Table 4 Standard deviation of random effects in uric acid level analysis

	Random Effect	Standard Deviation
	Intercept	1.57
	Residual	1.96

Table 5 Covariance parameters and correlation coefficients

Covariance parameter	Correlation with the second visit	Correlation with the third visit	Correlation with the fourth visit
Second visit	–	–	–
Third visit	0.323	–	–
Fourth visit	0.199	0.207	–
Fifth visit and later	0.227	0.236	0.273

A post-hoc analysis indicated that the significance of changes in uric acid levels was apparent as early as the second visit, with an estimated reduction of 1.4 mg/dL (SE=0.0489, p-value<0.0001: Supplementary Table 2).

Predictive modelling using machine learning

Following data processing, the variables in the investigated dataset that demonstrate completeness and quality include: gender, age, uric acid levels, and the category indicating whether a patient is considered to have hyperuricemia. These variables were then subjected to a Support Vector Machine (SVM) algorithm to construct a predictive model. SVM was selected for its proven ability to effectively manage the complexity inherent in datasets with a limited number of features but requiring nuanced differentiation. The data processing successfully transformed the reduction of uric acid in the patient cohort into a binary classification, then coded this into a separate feature as: 'response'; including 'good', which refers to a reduction of uric acid in the subsequent visit, and 'poor', which refers to the incremental changes in uric acid levels.

The SVM model, with a linear kernel, demonstrated an overall accuracy of 66.49% (CI: 64.65%–68.29%). The obtained results represent a significant improvement over the No Information Rate (NIR) of 59.75%, wherein the model predicts the majority class without learning or gaining information from the features. Additionally, the model's predictive capability is statistically significant (p-value<0.0001). Further analysis of the model's performance revealed a sensitivity (true positive rate) of 59.27% and a specificity (true negative rate) of 71.36%. The positive predictive value (PPV) was calculated to be 58.23%, while the negative predictive value (NPV) was 72.23%. The kappa statistic, a measure of agreement beyond chance, was found to be 0.3053, indicating a fair level of agreement. Additionally, McNemar's Test for the symmetry of the model's misclassification matrix produced a p-value of 0.5451, suggesting no significant bias in the

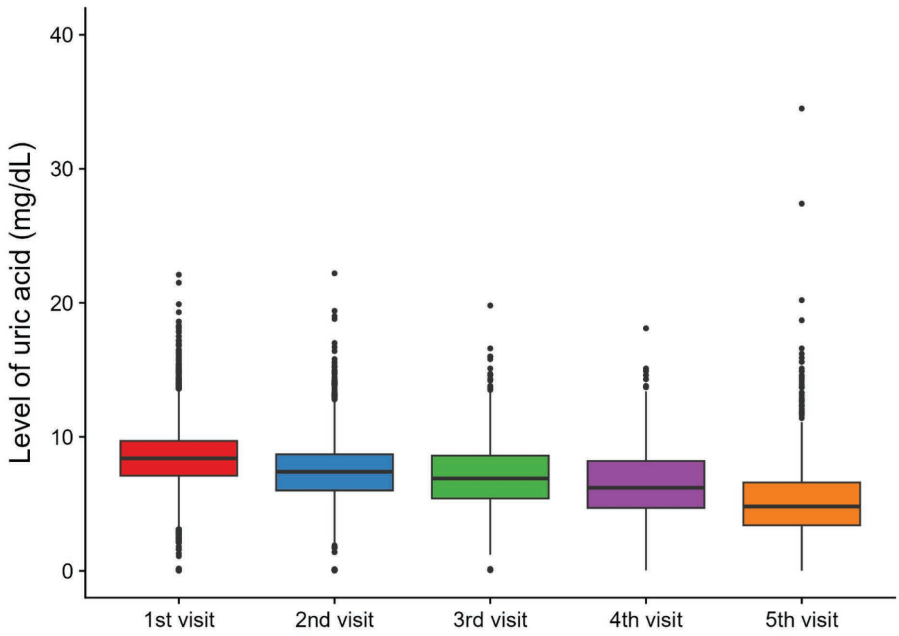


Figure 1 The overall uric acid level decreased across five visits. Even when neglecting the subgroup depending on age, the trend of reduction shows a marginal decrease across subsequent visits

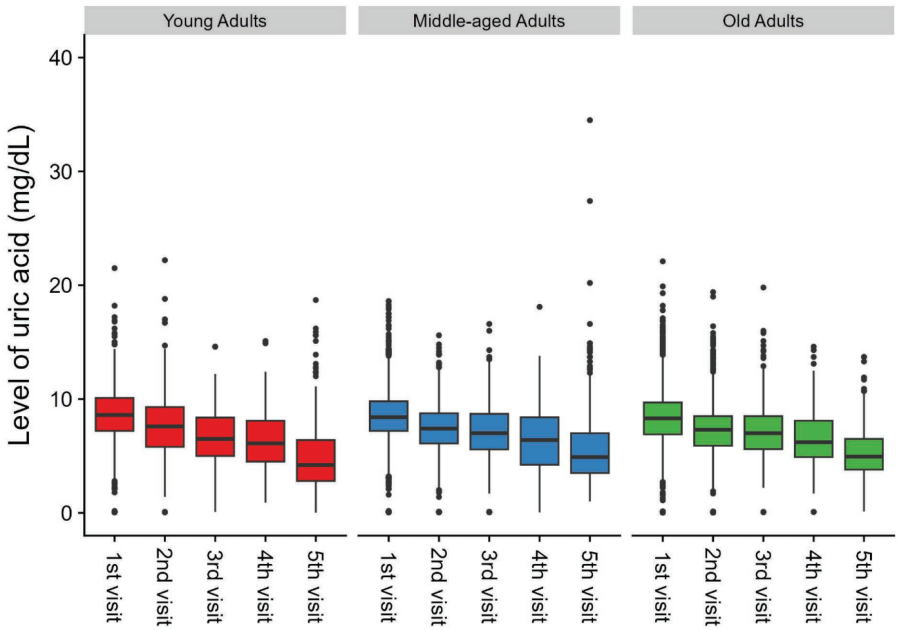


Figure 2 Uric acid levels of patients across visits among different age groups. The trend of reduction in young, middle-aged, and older adults has exhibited a successive decrease across subsequent hospital visits

prediction of 'good' compared to 'poor' response. The balanced accuracy of the model, a metric that accounts for both sensitivity and specificity in imbalanced datasets, was 65.31%.

Safety profile of allopurinol

Further analyses were conducted to investigate the safety profile of allopurinol. During the data mining procedure, the overall ADR report data was limited due to unstructured and missing data. However, through both manual and automated mining approaches, the ADR data was successfully extracted, resulting in a dataset of 385 records. The majority of ADRs, more than 50%, were rashes, as shown in (Figure 3), while severe cutaneous reactions, such as SJS and DRESS, ranked as the third and fifth most commonly reported ADRs observed in patients exposed to allopurinol. Subsequently, patients who had undergone HLA screening tests were further analyzed for subgroup analysis, as it is well-established that the presence of *HLA-B*58:01* is associated with the occurrence of severe skin reactions. Practically, in the previous clinical practice setting, most patients tend to develop particular forms of ADR before undergoing an HLA screening test to confirm the etiology of the ADR, whether this is correlated to the presence of *HLA-B* genes or not. As such, only 185 patients were screened for *HLA-B*58:01* and included in this descriptive analysis. Of the 185 patients, only 15 patients (0.08%) reported experiencing ADR derived from allopurinol treatment (Supplementary Table 1).

Discussion

Recently, the analysis of big data has been at the center of attention as one of the prominent contemporary research approaches. Notably, within the research field of healthcare, the application of big data analysis aims to extract insights from real-world data generated on a daily

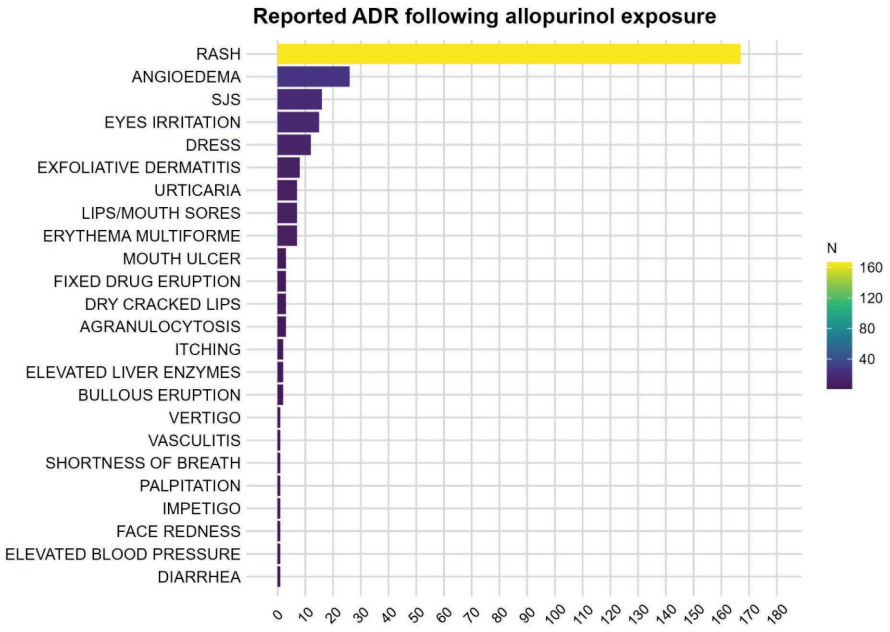


Figure 3 Descriptive analysis of reported adverse reactions following allopurinol exposure. The events are organized by frequency, and the bar graph employs varying shades to depict the events in accordance with their frequency

basis[20]. This real-world data serves as compelling evidence, showcasing practical scenarios where real-life issues become evident, leading to tangible changes. On the other hand, real-world data has been acquired at a cost, it often presents a challenge due to its substantial unstructured nature, rendering it difficult to extract pragmatic information. The findings described in this study provides evidence to support the efficacy of the clinical practice guidelines that has been implemented for the administration of allopurinol in Songklanagarind Hospital; as it effectively and successfully reduced uric acid levels. In support of this conclusion, Supplementary Figure 1 provides a comprehensive overview of how allopurinol has been prescribed in the investigated hospital over the last decade, wherein most patients receive 100 mg of allopurinol. For a detailed examination of allopurinol prescription trends and their implications, please refer to Supplementary Figure 1.

This present study employed a multi-faceted data mining approach of the hospitals-level data. It incorporated techniques such as regular expression and text mining in tandem with visualization and feature extraction. These methodologies were orchestrated iteratively to ensure a cohesive analytical process. Consequently, the inherent, unstructured nature of the initial data was gradually transmuted into a coherent and understandable output[21]. Within this context, a decade-long dataset pertaining to allopurinol was subjected to analysis, offering diverse insights ranging from patient attributes to the intricacies of safety profiles. Initially, a proof-of-concept assessment was conducted to ascertain the efficacy of the existing allopurinol regimen in effectively managing uric acid levels. Hence, over the course of multiple visits with a regularly scheduled follow-up interval of approximately one month, a successive decrease in the level of uric acid was observed, leading to a statistically significant therapeutic outcome. In general, the prescribed dosage for allopurinol to achieve the intended uric acid-lowering effect in patients with hyperuricemia is 300 mg per day. Nevertheless, owing

to the frequent occurrence of severe skin reactions; especially among individuals of Asian ethnicity, the practical guidelines in Thailand have undergone gradual adaptation over the past decade. A 'start slow, go slow' approach may be pragmatically applied to reduce the risk of severe skin reactions and serves as a concrete guideline[11, 22]. Regrettably, while this study implicitly aims to elucidate the reduced dose of allopurinol administered at Songklanagarind Hospital, limitations in data recruitment through internal infrastructure prevented a thorough subgroup analysis of daily dosing among patients. This, therefore, hindered the legitimacy of support for policy changes. Nevertheless, over the past decade, the efficacy of allopurinol has been undeniably demonstrated: it continues to be effective. Moreover, a machine learning approach employing an SVM with a linear kernel has been successfully developed, showing moderate accuracy in predicting patterns of uric acid reduction response. With this model, introducing new, unknown data that matches the features described in the model can, at the very least, accurately predict the response with up to a 66% accuracy rate, among other properties detailed in the results section. The decision to use a linear kernel over a non-linear one was based on the similarity in predictive results. Although a non-linear kernel might provide greater flexibility in navigating the complex relationships among variables, this added complexity can challenge the simplification of the model's predictions. Such simplification is often paramount in clinical settings, wherein a model's interpretability significantly influences decision-making processes.

Thailand has recently implemented a health policy requiring *HLA-B*58:01* screening before the initiation of allopurinol treatment, which should substantially decrease the probability of encountering multiple cases of drug allergies related to the medication. Additionally, if a thorough investigation confirms the efficacy of the reduced starting dose regimen, this approach will continue to uphold the well-established efficacy and safety of allopurinol.

The strength of this study lies in its meticulous organization of extensive real-world data collected over ten years. Nevertheless, due to the incorporation of historical and contemporary data, certain facets of analysis remained unfeasible. For instance, since the *HLA-B*58:01* records have only been gathered over the past 5 years (from 2017 onwards), the limited number of cases made it impractical to develop a predictive model. Significantly, this study is limited by missing data, primarily due to constraints in extracting information from the data in Hospital Information System, which may lead to the underestimation of certain paradigms. Furthermore, the present dataset has not taken into account concurrent medications and other traditional remedies that patients might have used between hospital visits. Moreover, pharmacovigilance in Thailand still struggles to maintain seamless reporting of data, leading to the underestimation of ADRs associated with certain medications, including in the case of allopurinol. Hence, further studies are required to address several of these questioning aspects in order to correctly extrapolate to the practical situation.

Conclusion

The standard allopurinol treatment approach demonstrated favorable treatment outcomes and a reduced likelihood of causing serious ADRs. A monthly follow-up schedule would be the optimal duration to evaluate treatment progress and enhance patient adherence to the prescribed regimen.

Acknowledgement

This work was supported by the Faculty of Science, Prince of Songkla University, Grant number SCI6504132S. The authors acknowledge all clinicians, nurses, pharmacists, and healthcare workers who worked for the care of all patients in this study.

Author contribution

WR: study design, data collection, data analysis, manuscript writing, manuscript editing, PV: study design, data analysis, manuscript editing, KJ: study design, data collection, manuscript editing.

Conflict of interest

The authors declare no conflict of interest.

References

1. Shagroni T, Cazares AR, Kim JA, Furst DE. Nonsteroidal anti-inflammatory drugs, disease-modifying antirheumatic drugs, nonopioid analgesics, & drugs used in gout. In: Katzung BG, Vanderah TW, editors. Basic & Clinical Pharmacology, 15 ed. New York: McGraw-Hill; 2021.
2. Bardin T, Richette P. The role of febuxostat in gout. Curr Opin Rheumatol 2019;31:152-8.
3. FitzGerald JD, Dalbeth N, Mikuls T, Brignardello-Petersen R, Guyatt G, Abeles AM, et al. 2020 American College of Rheumatology Guideline for the Management of Gout. Arthritis Care Res (Hoboken) 2020;72:744-60.
4. White WB, Saag KG, Becker MA, Borer JS, Gorelick PB, Whelton A, et al. Cardiovascular safety of febuxostat or allopurinol in patients with gout. N Engl J Med 2018;378:1200-10.
5. Becker MA, Schumacher HR Jr, Wortmann RL, MacDonald PA, Eustace D, Palo WA, et al. Febuxostat compared with allopurinol in patients with hyperuricemia and gout. N Engl J Med 2005;353:2450-61.
6. Yu KH, Yu CY, Fang YF. Diagnostic utility of HLA-B*5801 screening in severe allopurinol hypersensitivity syndrome: an updated systematic review and meta-analysis. Int J Rheum Dis 2017;20:1057-71.
7. Dean L, Kane M. Allopurinol Therapy and HLA-B*58:01 Genotype. In: Pratt VM, Scott SA, Pirmohamed M, Esquivel B, Kattman BL, Malheiro AJ, editors. Medical Genetics Summaries. Bethesda (MD): National Center for Biotechnology Information (US); 2012.
8. Tassaneeyakul W, Jantararoungtong T, Chen P, Lin P-Y, Tiamkao S, Khunarkornsiri U, et al. Strong association between *HLA-B*5801* and allopurinol-induced Stevens-Johnson syndrome and toxic epidermal necrolysis in a Thai population. Pharmacogenet Genomics 2009;19:704-9.
9. Saksit N, Tassaneeyakul W, Nakkam N, Konyoung P, Khunarkornsiri U, Chumworathayi P, et al. Risk factors of allopurinol-induced severe cutaneous adverse reactions in a Thai population. Pharmacogenet Genomics 2017;27:255-63.
10. Kloypan C, Koomdee N, Satapornpong P, Tempark T, Biswas M, Sukasem C. A Comprehensive review of HLA and severe cutaneous adverse drug reactions: implication for clinical pharmacogenomics and precision medicine. Pharmaceuticals (Basel) 2021;14.
11. Richette P, Doherty M, Pascual E, Barskova V, Becce F, Castañeda-Sanabria J, et al. 2016 updated EULAR evidence-based recommendations for the management of gout. Ann Rheum Dis 2017;76:29-42.
12. Kutanan W, Kampuansai J, Changmai P, Flegontov P, Schröder R, Macholdt E, et al. Contrasting maternal and paternal genetic variation of hunter-gatherer groups in Thailand. Scientific Reports 2018;8:1536.
13. Dancause KN, Chan CW, Arunotai NH, Lum JK. Origins of the Moken Sea Gypsies inferred from mitochondrial hypervariable region and whole genome sequences. J Hum Genet 2009;54:86-93.
14. Deshpande D, Chhugani K, Chang Y, Karlsberg A, Loeffler C, Zhang J, et al. RNA-seq data science: from raw data to effective interpretation. Front Genet 2023;14:997383.
15. Hayes CJ, Cucciare MA, Martin BC, Hudson TJ, Bush K, Lo-Ciganic W, et al. Using data science to improve outcomes for persons with opioid use disorder. Subst Abus 2022;43:956-63.
16. Slade E, Dwoskin LP, Zhang GQ, Talbert JC, Chen J, Freeman PR, et al. Integrating data science into the translational science research spectrum: a substance use disorder case study. J Clin Transl Sci 2020;5:e29.
17. Goulooze SC, Zwep LB, Vogt JE, Krekels EHJ, Hankemeier T, van den Anker JN, et al. Beyond the Randomized Clinical Trial: Innovative Data Science to Close the Pediatric Evidence Gap. Clin Pharmacol Ther 2020;107:786-95.
18. Ehsani-Moghaddam B, Martin K, Queenan JA. Data quality in healthcare: a report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data. Health Inf Manag 2021;50:88-92.
19. Rigobon AV, Birtwhistle R, Khan S, Barber D, Biro S, Morkem R, et al. Adult obesity prevalence in primary care users: an

exploration using Canadian primary care sentinel surveillance network (CPCSSN) data. Can J Public Health 2015;106:e283-9.
20. Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. BMC Med Res Methodol 2022;22:287.
21. Lin WL, Liang YC, Chung KH, Chen PH, Chang YC. Using text mining and data visualization approaches for investigating mental illness from the perspective of traditional Chinese medicine. Medicina (Kaunas) 2023;59.
22. Stamp LK, Taylor WJ, Jones PB, Dockerty JL, Drake J, Frampton C, et al. Starting dose is a risk factor for allopurinol hypersensitivity syndrome: a proposed safe starting dose of allopurinol. Arthritis Rheum 2012;64:2529-36.

Supplementary Table 1 Subgroup analysis of cases with HLA screening results

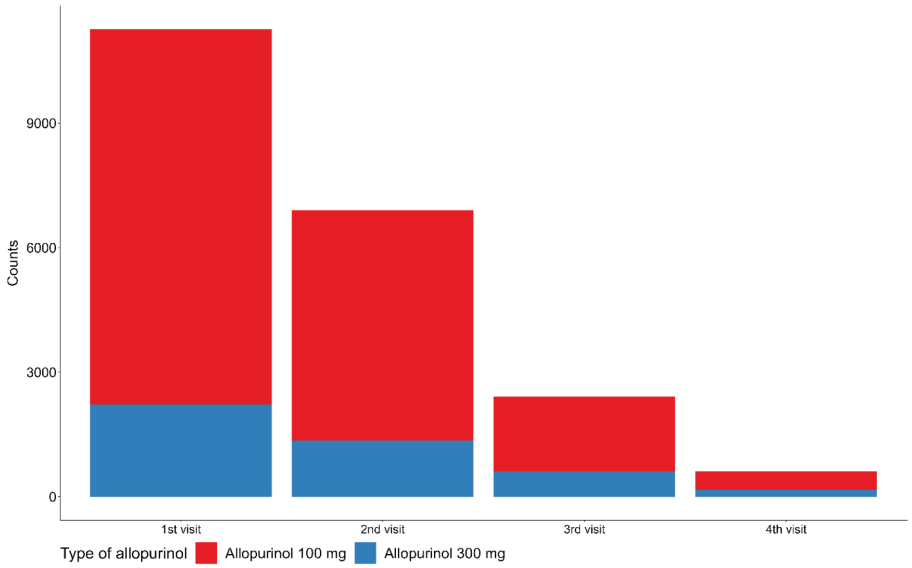
	Young Adults	Middle-aged Adults		Old Adults		Overall	
	Male	Female	Male	Female	Male	Female	Male
	(N=13)	(N=5)	(N=38)	(N=30)	(N=99)	(N=35)	(N=150)
Presence of adverse events							
No ADR	13 (100%)	5 (100%)	33 (86.8%)	28 (93.3%)	91 (91.9%)	33 (94.3%)	137 (91.3%)
Yes ADR	0 (0%)	0 (0%)	5 (13.2%)	2 (6.7%)	8 (8.1%)	2 (5.7%)	13 (8.7%)
Primary ADR							
None	13 (100%)	5 (100%)	33 (86.8%)	28 (93.3%)	91 (91.9%)	33 (94.3%)	137 (91.3%)
DRESS	0 (0%)	0 (0%)	1 (2.6%)	0 (0%)	1 (1.0%)	0 (0%)	2 (1.3%)
Rash	0 (0%)	0 (0%)	4 (10.6%)	0 (0%)	6 (6.1%)	0 (0%)	10 (6.7%)
Shortness of breath	0 (0%)	0 (0%)	0 (0%)	1 (3.3%)	0 (0%)	1 (2.9%)	0 (0%)
SJS	0 (0%)	0 (0%)	0 (0%)	1 (3.3%)	0 (0%)	1 (2.9%)	0 (0%)
Eyes irritation	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (1.0%)	0 (0%)	1 (0.7%)
HLA test results							
Negative	12 (92.3%)	5 (100%)	37 (97.4%)	27 (90.0%)	93 (93.9%)	32 (91.4%)	142 (94.7%)
Positive	1 (7.7%)	0 (0%)	1 (2.6%)	3 (10.0%)	6 (6.1%)	3 (8.6%)	8 (5.3%)

N=number of patients, ADR=adverse drug reaction, DRESS=drug reaction with eosinophilia and systemic symptoms, HLA=human leukocyte antigens

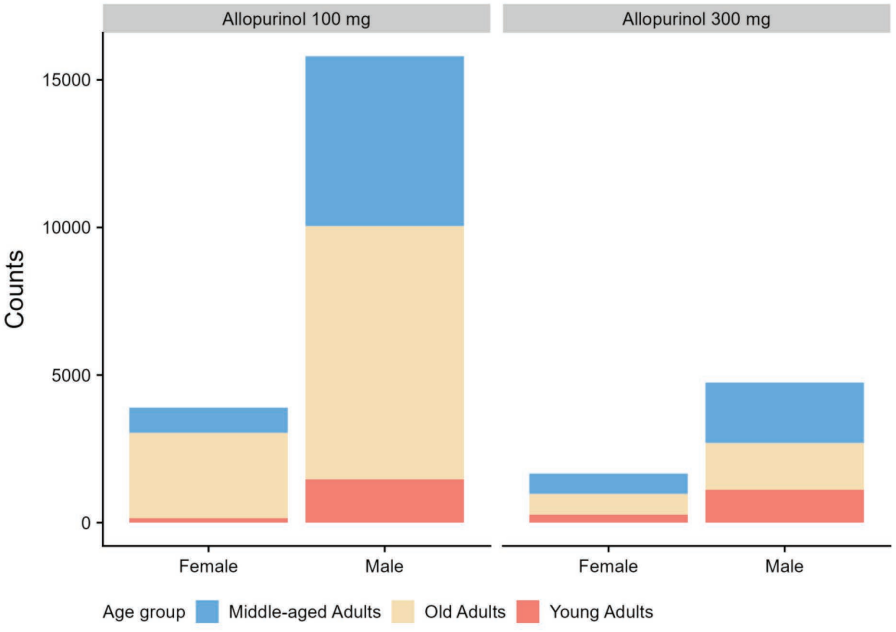
Supplementary Table 2 Post-hoc analysis following the analysis of deviance in a linear mixed model

Contrast	Estimate	SE	t-ratio	p-value
Visit 1 – Visit 2	1.429	0.0489	29.242	<0.001
Visit 1 – Visit 3	1.744	0.0756	23.060	<0.001
Visit 1 – Visit 4	1.988	0.1230	16.158	<0.001
Visit 1 – Visit 5 and later	3.009	0.1078	27.922	<0.001
Visit 2 – Visit 3	0.341	0.0756	4.156	0.0003
Visit 2 – Visit 4	0.558	0.1230	4.538	0.0001
Visit 2 – Visit 5 and later	1.580	0.1078	14.658	<0.001
Visit 3 – Visit 4	0.244	0.1304	1.872	0.3328
Visit 3 – Visit 5 and later	1.265	0.1161	10.868	<0.001
Visit 4 – Visit 5 and later	1.021	0.1397	7.312	<0.001

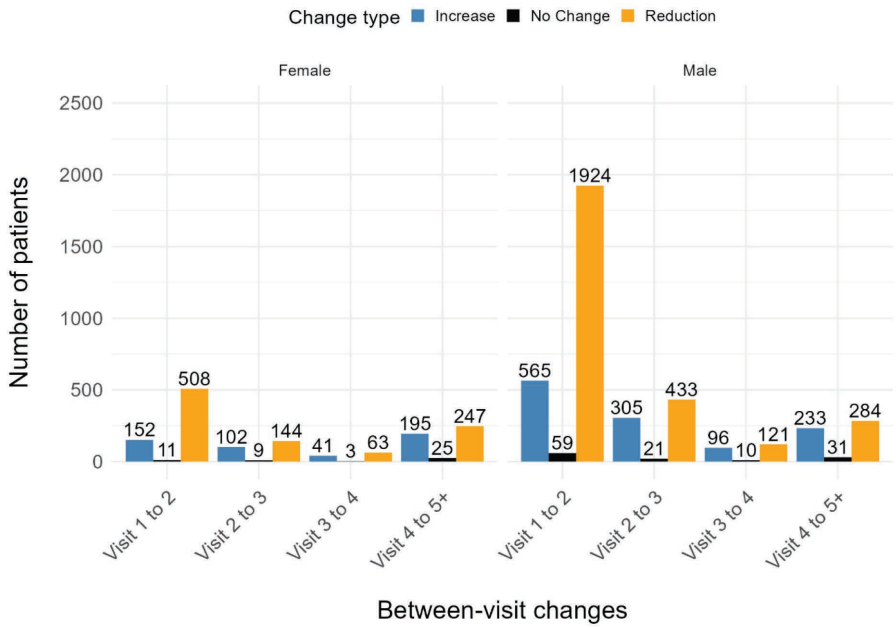
P-value adjustment: Tukey method for comparing a family of 5 estimates. Confidence level used: 0.95
SE=standard error



Supplementary Figure 1 Prescribed allopurinol dosage in each visit; from 2010 and 2022. The figure shows only four visits: as the fifth visit has a very limited number of records. The y-axis represents the number of records extracted from the electronic medical record (EMR)



Supplementary Figure 2 Prescribed allopurinol dosage for each gender and age group between 2010 and 2022 The y-axis represents the number of records extracted from the electronic medical record (EMR)



Supplementary Figure 3 Changes in uric acid levels between visits following allopurinol treatment; categorized by gender. Patients are divided into three groups: those exhibiting a reduction in uric acid levels, those with an increase, and those with no change. The numbers above each bar indicate the patient count within each category