Received: 22 October 2020 | Accepted: 24 October 2020

DOI: 10.1002/prp2.687

ORIGINAL ARTICLE

Classification of drug use patterns

Christiaan H. Righolt | Geng Zhang | Salaheddin M. Mahmud

Vaccine and Drug Evaluation Centre, Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, Canada

Correspondence
Christiaan H. Righolt, Vaccine and Drug Evaluation Centre, Department of Community Health Sciences, University of Manitoba, 337 - 750 McDermot Avenue, Winnipeg, Manitoba, R3E 0T5, Canada.
Email: Christiaan.Righolt@umanitoba.ca

Funding information
This work was supported by a Research Manitoba Research Cluster Grant (# 761090863). The opinions presented in the report do not necessarily reflect those of the funder. SMM's work is supported, in part, by funding from the Canada Research Chair Program.

Abstract

Characterizing long-term prescription data is challenging due to the time-varying nature of drug use. Conventional approaches summarize time-varying data into categorical variables based on simple measures, such as cumulative dose, while ignoring patterns of use. The loss of information can lead to misclassification and biased estimates of the exposure-outcome association. We introduce a classification method to characterize longitudinal prescription data with an unsupervised machine learning algorithm. We used administrative databases covering virtually all 1.3 million residents of Manitoba and explicitly designed features to describe the average dose, proportion of days covered (PDC), dose change, and dose variability, and clustered the resulting feature space using K-means clustering. We applied this method to metformin use in diabetes patients. We identified 27,786 metformin users and showed that the feature distributions of their metformin use are stable for varying the lengths of follow-up and that these distributions have clear interpretations. We found six distinct metformin user groups: patients with intermittent use, decreasing dose, increasing dose, high dose, and two medium dose groups (one with stable dose and one with highly variable use). Patients in the varying and decreasing dose groups had a higher chance of progression of diabetes than other patients. The method presented in this paper allows for characterization of drug use into distinct and clinically relevant groups in a way that cannot be obtained from merely classifying use by quantiles of overall use.

KEYWORDS

Clustering, drug exposure, K-means, machine learning, pharmacoepidemiology

1 | INTRODUCTION

Exposure to prescription drugs needs to be characterized to investigate the association between drug exposure and specific health outcomes. When detailed longitudinal exposure data are available, characteristics of the data, for example, time since exposure and intensity of exposure, vary over time and are harder to interpret than short-term exposures, posing challenges in studying the exposure-outcome association.[1,2] Conventional approaches to characterizing long-term drug exposure summarize use by collapsing longitudinal data to single measures, such as ever-use, average dose of use, and duration of use. Classification

Abbreviations: DDD, defined daily dose; DPIN, Drug Program Information Network; HAD, Hospital Abstracts Database; ICD-10-CA, International Classification of Diseases, Tenth Revision, Canadian Edition; ICD-9-CM, International Classification of Diseases, Ninth Revision, Clinical Modification; MH, Manitoba Health; MHPR, Manitoba Health Population Registry; MSD, Medical Services Database; PDC, proportion of days covered; PHIN, personal health identification number.

Christiaan H. Righolt and Geng Zhang: Joint first authors.

using these simple measures may not always account for the true differences between patient groups. For example, patients using stable, increasing, decreasing, or frequently varying doses can have the same mean dose. The loss of information can cause misclassification of the drug exposure and lead to biased estimates of the association.[3–6]

The cumulative and mean dose are the uniformly weighted sum and average of the dose over time. Several studies have extended this approach and summarized the cumulative dose by nontrivial weights.[7–13] The weights are estimated from non- or quasi-parametric methods. These methods still aim to summarize time-varying exposure as a single-valued measure.

The prescription data can be used directly in regression analysis through interrupted time series methods. Because there may be patient groups with prescription patterns of clinical relevance, group-based trajectory modeling (GBTM) was introduced to uncover mean group patterns.[14–16] Original GBTM assumes a polynomial trend of the time series and is unable to describe complex patterns, such as periodic use (possibly because of seasonal effects) or nonadherent/occasional use (in which the dose variations do not approximate low-order polynomials), which may generate patterns that are not supported by the data.[17] Fitting the group patterns with B-splines relaxes the constrains on the functional form of the hypothetical patterns.[17,18] Sometimes patients have irregular/varying prescription patterns, and this pattern may be "averaged out" when patterns are averaged for a whole group in GBTM. Since the fitting process involves maximizing a likelihood function, GBTM may converge slowly as the number of time points for each time series or the total number of time series grows, or it may even fail to converge when using higher orders of polynomial functions. [19–21]

Recent advances in machine learning provide more statistical tools to understand and characterize time series. Methods such as recurrent neural networks (RNN)[22–25] and long- and short-term memory auto-encoders followed by K-means clustering[26] have been used in various applications. RNN auto-encoder networks require no human input for feature extraction (the process of deriving variables) and require no assumptions about the input time series. Due to the nonlinearity of RNN networks, it is almost impossible to interpret the derived features (derived variables, which may not have a human interpretation). In medical research, it is important to translate research results into clear clinical terms to benefit patient care.

The success and limitations of existing methods inspired us to characterize time-varying exposure based on interpretable features (derived variables) of drug use. In this article, we demonstrate a method to classify drug use patterns into clinically relevant groups. Without assuming any specific form of the exposure patterns, we explicitly define a list of features summarizing individual-level prescription data. We first introduce our use case (which we use to exemplify this method). Then, we describe our classification method, which includes a description and rationale for the features used to describe drug use patterns and the classification of the resulting feature space (a multidimensional space in which each feature/variable is one dimension/axis). Finally, we investigate the results for our use case before discussing the implications of this method.

2 | MATERIALS AND METHODS

2.1 | Data sources and variables

Manitoba Health (MH) is the publicly funded health insurance agency providing comprehensive health insurance, including coverage for hospital and outpatient physician services, to the province's 1.3 million residents. Coverage is universal, with no eligibility distinction based on age or income, and participation rates are very high (>99%).[27] MH maintains several centralized, administrative electronic databases that are linkable using a unique personal health identification number (PHIN). The completeness and accuracy of these databases are well established.[28,29]

The Drug Program Information Network (DPIN), in operation since 1995, records all prescription drugs dispensed to Manitoba residents electronically at the "point-of-sale", including most personal care home residents.[30] The DPIN database captures data from pharmacy claims for formulary drugs dispensed to all Manitobans even those without prescription drug insurance. Since 1971, the Hospital Abstracts Database (HAD) recorded virtually all services provided by hospitals in the province, including admissions and day surgeries.[28] The data collected comprise demographic as well as diagnosis and treatment information including primary diagnosis and service or procedure codes, coded using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) before April, 2004, and the ICD-10-CA (Canadian adaptation of the ICD-10) and the Canadian Classification of Health Interventions (CCI) afterwards. The Medical Services Database (MSD), also in operation since 1971, collects similar information, based on physician fee-for-service or shadow billing, on services provided by physicians in offices, hospitals, and outpatient departments across the province.[28]

We picked metformin use to exemplify this method, because it is a common long-term first-line medication in type 2 diabetes. We identified all persons covered by MH and diagnosed with diabetes (Table S1) during 1995-2017 and assessed their metformin use from DPIN from their first filled prescription until March 2017 in successive 90-day periods (and limited follow-up to 1, 2, 5, and 10 years after the first prescription in analyses below).

We obtained income quintile and residence through the MHPR linked to the Canadian Census, we defined diabetes progression as insulin use (from DPIN) or diabetic complications (from the HAD and MSD), and obtained the number of physician visits (excluding repeat visits on the same day) from the MSD.

We used SAS 9.4 (SAS Institute) for general data preparation and cleaning and R 3.5.1 (R Foundation[31]) for clustering. This study was approved by the University of Manitoba Research Ethics Board and by MH's Health Information Privacy Committee.

2.2 | Drug use classification

2.2.1 | Measurements

Both drug prescription histories and individual prescriptions vary in length. We standardized raw data into τ-day episodes starting at first use. The start dates of refills were shifted to start after the previous refill would be used. We used $\tau = 90$, which is a typical prescription length, other choices could include 60 or 120 days (the choice of τ does not affect the principle of the method, but may alter the results and should be evaluated for each application). For each person i and each period j, we identified: 1) the number of episodes, N_i, for the person, 2) the number of days, t_{ij}, the drug was prescribed during the episode, 3) the length of follow-up in days, f_{ij}(it is possible that follow-up ends before the τ-day episode), and 4) the total dose, d_{ij}, during the episode (calculated as the defined daily dose [DDD], "the assumed average maintenance dose per day for a drug used for its main indication in adults"[32]).

2.2.2 | Features

The original time series form a high-dimensional space. To reduce this dimensionality, we created derived features (derived variables), which prevents overfitting and helps interpretation of the subsequent classification. We require that features are clinically interpretable and that they capture the characteristics of patterns exhibited in the study data. Because the total length of follow-up is not equal for all patients and because we want to scale our algorithm to different durations, features should be independent of N_i (the total follow-up for each person). A dose-response can have multiple forms for drug exposures; it could be the dose itself, the relative duration of use, the dose change (increasing/decreasing), and dose variability of use.

Length of follow-up

The length of follow-up, F_i, is:

$$F_i = \sum_{j=1}^{N_i} f_{ij} \tag{1}$$

for each person. We limited/fixed follow-up in our use case to simplify the interpretation of our results.

Average dose

The average dose during use D_i, is described as:

$$D_i = \frac{\sum_{j=1}^{N_i} d_{ij}}{\sum_{j=1}^{N_i} t_{ij}} \tag{2}$$

This is independent of the choice of τ (the length of each episode). D_i is the average dose during use, not the average during follow-up, because time without use is represented by the proportion of days covered (PDC) (see below).

Proportion of days covered (PDC)

The PDC is the proportion of follow-up a patient received the specific drug, after shifting prescriptions that were refilled early to account for potential drug stockpiling. The PDC for each person is:

$$PDC_i = \frac{\sum_{j=1}^{N_i} t_{ij}}{\sum_{j=1}^{N_i} f_{ij}} \tag{3}$$

Although it is not known whether a patient actually took the medication, the PDC will typically be lower for nonadherent patients, because they do not need to refill their medication in time (because they do not run out in time). The PDC is also independent of τ.

Dose change

For some exposure-outcome relations, it matters whether the dose was stable/constant or changed over time. We define the overall trend, T_i, by normalizing the linear least squares slope with the total number of episodes N_i:

$$T_i = \frac{12N_i}{N_i^2 - 1}\mathrm{cov}(\bar{\bar{d}}_i, \bar{p}_i) \tag{4}$$

where $\mathrm{cov}(\bar{\bar{d}}_i, \bar{p}_i)$ is the covariance between the average daily dose vector $\bar{\bar{d}}_i = (d_{i1}/t_{i1}, d_{i2}/t_{i2}, d_{i3}/t_{i3}, \ldots, d_{iN_i}/t_{iN_i})$ and episode vector and $\bar{p}_i = (p_{i1}, p_{i2}, p_{i3}, \ldots, p_{iN_i})$ See appendix for derivation.

Dose variability

Another important aspect of drug use patterns is the extent to which the dose fluctuates around the overall trend. Drug titration is not necessarily a linear process and some drugs are only prescribed occasionally, making the linear fit close to a constant dose. The variability, V_i, describes these fluctuations in dose as:

$$V_i = \frac{1}{N_i - 2}\sum_{j=2}^{N_i - 1}\left|2\frac{d_{ij}}{t_{ij}} - \frac{d_{i(j-1)}}{t_{i(j-1)}} - \frac{d_{i(j+1)}}{t_{i(j+1)}}\right| \tag{5}$$

See appendix for rationale.

Feature distribution

We calculated these features for the prescription history of all metformin users in the study population for fixed follow-ups to examine the effect of N_i on the other features. The cumulative distribution function (CDF) of each feature is shown in Figure 1 for 1, 2, 5, and 10 years of follow-up time (N_i is 4, 8, 20, and 40, respectively). Although all features seem to converge to a long-term average, the PDC and dose variability have the smallest difference between 1- and 10-year follow-up. The average dose shows discontinuities in the CDF at typical prescription strengths of multiples of 500 mg (0.25 DDD), which means these patients use that specific average dose. A larger proportion of the dose change has $T_i = 0$ (a constant dose) for shorter follow-up. Metformin is a long-term medication (often used for years), so we selected the 5-year follow-up (20 episodes of 90 days) for clustering, because the prescription pattern is close to longer term averages at that point.

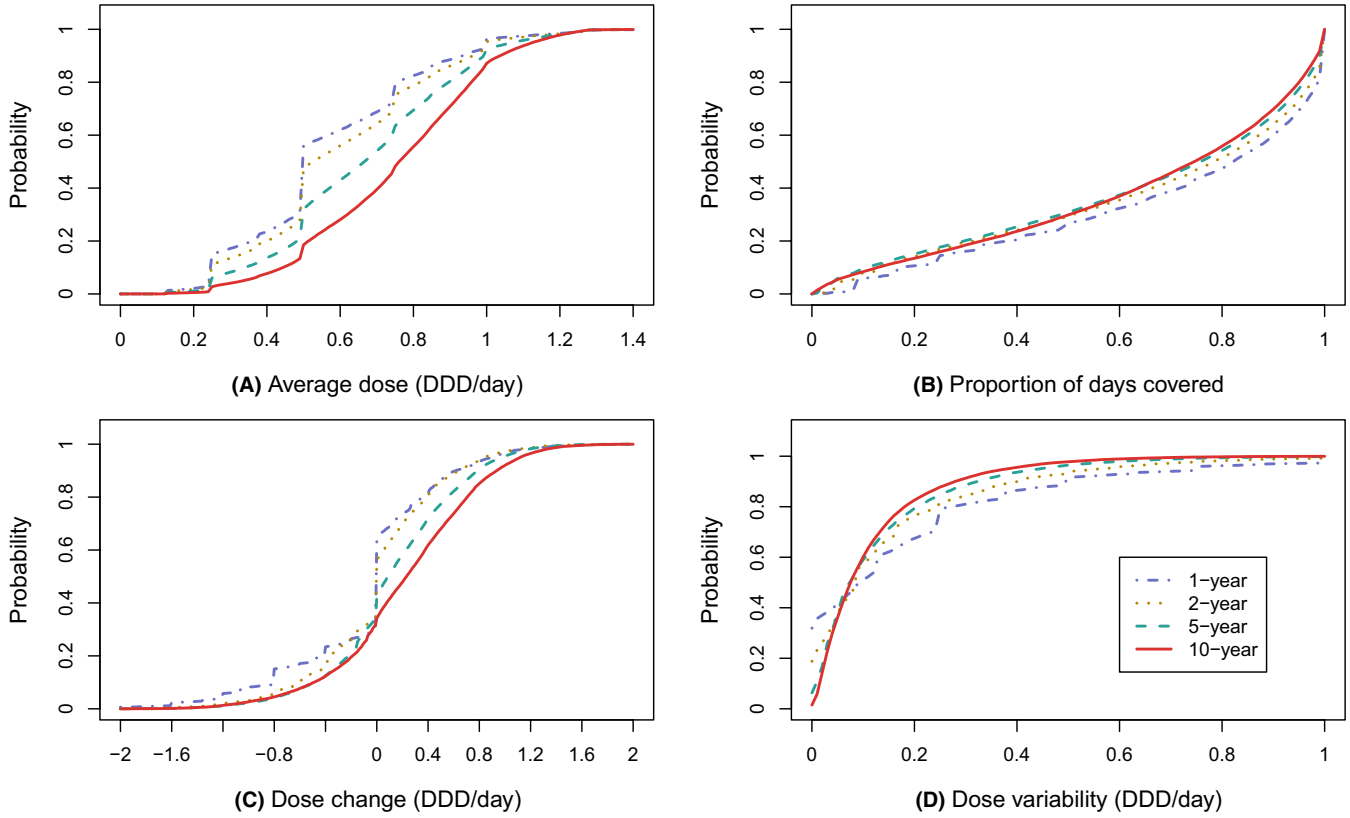

FIGURE 1 Cumulative distribution functions of drug use features according to the length of prescription data since the first dispensed metformin prescription. The defined daily dose (DDD) is 2 g for metformin

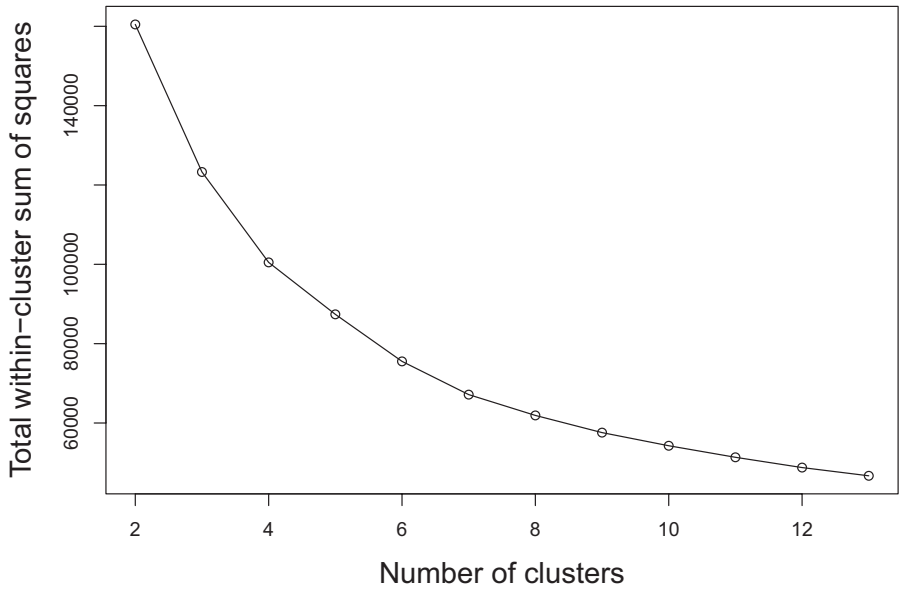

FIGURE 2 Within-cluster sum of squares according to the number of clusters

Clustering

The K-means clustering algorithm partitions data iteratively into a prespecified number (K) of clusters.[33] We used K-means clustering for its simplicity and convergence properties.[34] A commonly used technique to pick K is to plot the within-cluster sum of squares (WCSS) by different values of K and find the "elbow point".[35] The elbow point corresponds to K around 4-7 (Figure 2). Because there is no clear elbow point and because we want the classified groups/clusters to be clinically relevant, we evaluate K from 4-7 and pick the K for which K + 1 does not identify an additional use pattern (as interpreted by a human observer).

We averaged the drug use patterns within each cluster and plotted them for K=4, 5, 6 and 7 (Figure 3). Changing from *K*=4 to *K*=5, adds a cluster with decreasing dose. Changing to *K*=6 splits the group that ends with a high dose into an increasing dose group and a constant dose group. Changing to *K*=7 splits the increasing group into clusters with different rates of increase over time. Because there is little clinical difference between these groups, we opted for *K*=6. This classifies patients with intermittent use (solid red), increasing dose (dotted yellow), decreasing dose (long-dashed gray), high dose (short-/long-dashed green), and two medium dose groups (dash-dotted blue and short-dashed green). One of the medium dose groups (short-dashed green) has a stable dose, the other is highly variable (dash-dotted blue; these patients are either nonadherent or their prescription strengths fluctuate over time).

The same interpretation holds after projecting the feature space (a high-dimensional space in which each feature/variable is one dimension/axis) on a plane of each pair of features (Figure 4). Patients in the varying group differ predominantly in dose variability, but overlap with patients in terms of the other features. Patients with increasing and decreasing doses mostly stand out because of their dose change, although patients with decreasing use seem to have a lower PDC. Patients in the intermittent user and medium dose groups both have a relatively low average dose; the PDC is much lower for the intermittent user group (indicating more sporadic use), whereas the medium user group seems to use metformin for the majority of the time. Patients in the high-dose group use around 1 DDD/day (2 g/day) during most of the follow-up period.

3 | RESULTS

We identified 27,786 metformin users among diabetes patients and linked the classification results described above to hospital and medical service databases to characterize the resulting patient groups. The socio-economic and clinical characteristics of patients in different use groups vary (Table 1). The patient group with a varying dose consists of more persons younger than 45 (47.6%) and less persons 65 or older (8.3%) compared to other groups, especially the medium dose group with a similar average dose profile (16.1% for under 45 and 28.7% for 65+). Patients in the intermittent user, decreasing dose, and varying dose groups are relatively overrepresented in the lower income quintile (28.7%, 30.2%, and 33.5%) compared to the medium, increasing, and high-dose groups (21.1%, 24.0%, and 22.5%). Patients in the varying and decreasing dose groups more

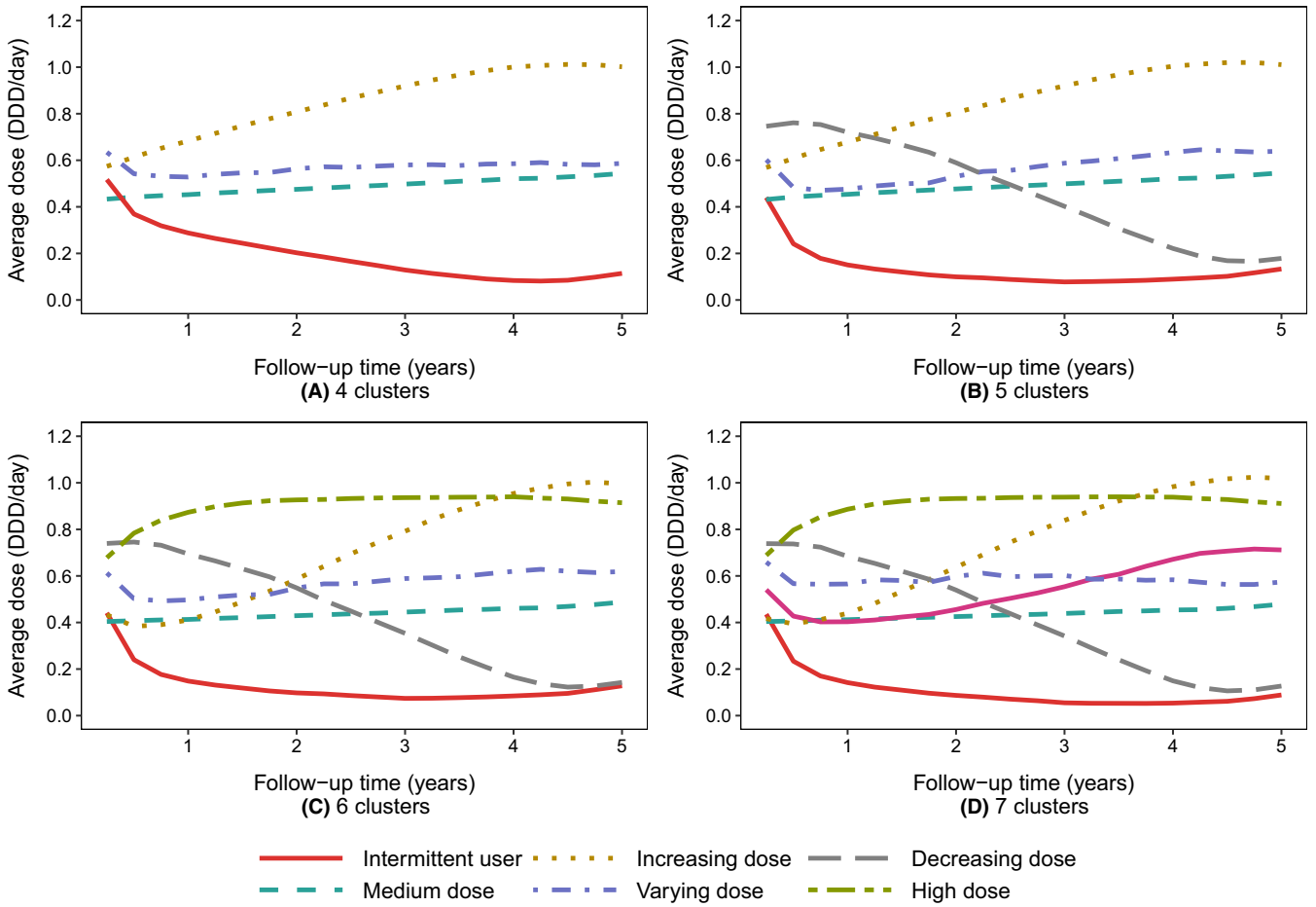

FIGURE 3 Average dose of the patients within each cluster over time according to the number of groups (K = 4 - 7) for K-means clustering. The defined daily dose (DDD) is 2 g for metformin

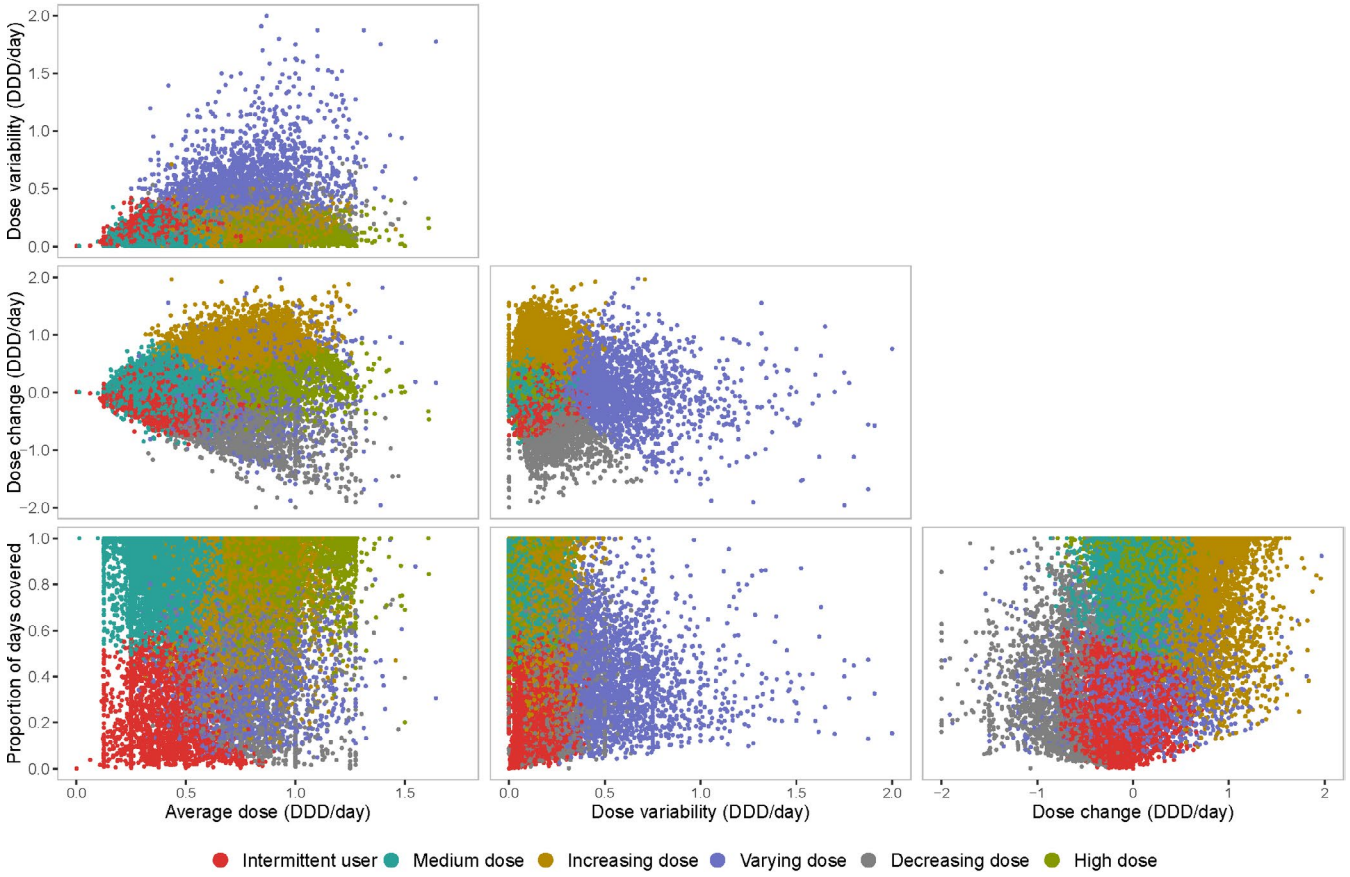

FIGURE 4 Scatter plots of drug use for each individual patient for each pair of features after classification into six clusters. The defined daily dose (DDD) is 2 g for metformin

frequently (17.8%) progress in their diabetes (defined as using insulin or having complications related to diabetes) than other patients, especially those in intermittent user and medium dose groups (9.6% and 10.4%).

4 | DISCUSSION

The role of metformin as a first-line treatment in type 2 diabetes is well known.[36,37] We found that a higher proportion of patients using a varying dose (possibly related to poor glycemic control) and patients using a higher dose progressed in their diabetes. Decreasing metformin use could mean that the medication is stopped, because poor glycemic control necessitates other treatment (patients in this group also have a higher chance of progression).[37] Patients with the same mean dose could have very different patterns of use (constant, increasing, decreasing, and varying), and we showed that this pattern of use is relevant as the clinical characteristics differ between these groups. These important differences would disappear if patients are classified by quantile based on overall dose. It is possible, however, that some of these differences are attributable to bias and confounding, including the healthy user effect and social determinants of health.

Our study did not include all diabetes patients in Manitoba, and our algorithm to detect diabetes only has a positive predictive value around 70% and a negative predictive value over 99%.[38] Because we are only interested in diabetes patients who use metformin, and because metformin refills require physician visits, we likely include the vast majority of metformin users in Manitoba. We used cross-validation (see below) to assess the impact of leaving out a portion of users on the results of the classification method.

In addition to the 5-year follow-up classification, presented in Figure 4, we used the same classification algorithm for 10 years of follow up and found similar results (data not shown). This means that one set of clusters derived from the K-means clustering algorithm might be used for patients with varying lengths of follow-up if this length is not included as a feature. To accommodate varying follow-up, the length of follow-up should be included as an additional input variable to the K-means algorithm. Instead of classifying a lifelong pattern, the follow-up period could also be broken in successive *M*-year periods for which each period is classified into a use group.

We did not assess how data quality affects this method, for example, how missing data or data entry errors affect the resulting groups. Because the results remain stable for different lengths of follow-up, we believe that the method is stable for deviation in the input data, but follow-up studies are required to understand the exact impact. There is uncertainty in this method; patients whose

TABLE 1 Number (percentage) of diabetes cases according to metformin use pattern groups with 5 years of follow-up according to certain socio-economic and clinical characteristics

	K-means groups					
	Intermittent user (N = 4475)	**Decreasing dose (N = 1953)**	**Medium dose (N = 9166)**	**Varying dose (N = 2422)**	**Increasing dose (N = 4581)**	**High dose (N = 5189)**
Gender						
Male	2,223 (49.7%)	1,016 (52.0%)	4,626 (50.5%)	1,303 (53.8%)	2,530 (55.2%)	2,971 (57.3%)
Female	2,252 (50.3%)	937 (48.0%)	4,540 (49.5%)	1,119 (46.2%)	2,051 (44.8%)	2,218 (42.7%)
Age at the diagnosis date of diabetes						
<45	1,488 (33.3%)	760 (38.9%)	1,480 (16.1%)	1,152 (47.6%)	1,140 (24.9%)	1,212 (23.4%)
45 - 54	1,154 (25.8%)	546 (28.0%)	2,457 (26.8%)	670 (27.7%)	1,599 (34.9%)	1,717 (33.1%)
55 - 64	940 (21.0%)	383 (19.6%)	2,600 (28.4%)	398 (16.4%)	1,206 (26.3%)	1,461 (28.2%)
65+	893 (20.0%)	264 (13.5%)	2,629 (28.7%)	202 (8.3%)	636 (13.9%)	799 (15.4%)
Income quintile						
Q1 (lowest)	1,286 (28.7%)	589 (30.2%)	1,930 (21.1%)	811 (33.5%)	1,101 (24.0%)	1,169 (22.5%)
Q2	975 (21.8%)	453 (23.2%)	1,982 (21.6%)	515 (21.3%)	1,021 (22.3%)	1,143 (22.0%)
Q3	825 (18.4%)	334 (17.1%)	1,886 (20.6%)	380 (15.7%)	884 (19.3%)	955 (18.4%)
Q4	765 (17.1%)	326 (16.7%)	1,852 (20.2%)	412 (17.0%)	851 (18.6%)	1,082 (20.9%)
Q5 (highest)	605 (13.5%)	235 (12.0%)	1,435 (15.7%)	290 (12.0%)	694 (15.1%)	797 (15.4%)
Unknown	19 (0.4%)	16 (0.8%)	81 (0.9%)	14 (0.6%)	30 (0.7%)	43 (0.8%)
Residence						
Rural	<1,956 (<43.7%)	<883 (<45.2%)	3,682 (40.2%)	<1,121 (<46.3%)	<1,902 (<41.5%)	2,250 (43.4%)
Urban	2,517 (56.2%)	1,067 (54.6%)	5,461 (59.6%)	1,301 (53.7%)	2,675 (58.4%)	2,919 (56.3%)
Unknown	<6 (<0.1%)	<6 (<0.3%)	23 (0.3%)	<6 (<0.2%)	<6 (<0.1%)	20 (0.4%)
Diabetes progression[a]	430 (9.6%)	347 (17.8%)	953 (10.4%)	431 (17.8%)	553 (12.1%)	762 (14.7%)
Insulin use	217 (4.8%)	210 (10.8%)	393 (4.3%)	289 (11.9%)	391 (8.5%)	472 (9.1%)
Diabetes complication	213 (4.8%)	138 (7.1%)	560 (6.1%)	142 (5.9%)	162 (3.5%)	290 (5.6%)
No. of physician visits during the 5-year period before diabetes diagnosis						
1 - 11	773 (17.3%)	421 (21.6%)	1,219 (13.3%)	477 (19.7%)	759 (16.6%)	991 (19.1%)
12 - 24	980 (21.9%)	423 (21.7%)	1,874 (20.4%)	611 (25.2%)	1,053 (23.0%)	1,178 (22.7%)
25 - 44	1,159 (25.9%)	500 (25.6%)	2,725 (29.7%)	646 (26.7%)	1,263 (27.6%)	1,382 (26.6%)
45+	1,563 (34.9%)	609 (31.2%)	3,348 (36.5%)	688 (28.4%)	1,506 (32.9%)	1,638 (31.6%)

[a]Insulin use or diabetic complications after initiating metformin.

drug use is right on the border between multiple user groups could be classified in different groups based on small fluctuations in the input data. We did not define uncertainty in the classification, for example, patients on the border between classes could fall in a different group depending on the input data. As an alternative, we performed cross-validation to assess the stability of the clustering process. Since there is no ground-truth for clustering, we use the cluster indices obtained from using all the input records as the "truth". We randomly divided the records into 10, 5, and 2 mutually excluded equal partitions. Each partition was used as the test set, while the complimentary data were used as the training set. The centroids obtained from clustering the training sets were applied to the test set. The assigned cluster indices were then compared to the "truth". 99.66%, 99.67%, and 99.34%, respectively, on average were clustered to the same group for 10-, 5-, and 2-fold cross-validations.

One of the major limitations of the K-means clustering algorithm is the requirement of a preset value for K, the number of clusters. The definition of a cluster is ill-defined, in extreme cases there may be as many clusters as the number of input time series. A variety of methods have been proposed to select K,[33] including information criterion-derived methods.[39] These data-driven methods lead to $K >> 6$ in most cases. We selected $K=6$, because little clinical information is gained by picking more clusters. Partitioning feature space into more clusters could result in overfitting the feature space on random fluctuations in drug use patterns. Other limitations of the K-means method include sensitivity to initial

values and outliers[33]; local optima can be generated for certain initial values. In our study, we initialized the centers with multiple random seeds[40,41] and obtained stable clustering results. For dealing with outliers, we excluded records with mean dose greater than 5.0 DDD during any τ-day episode. Outliers can also be excluded by the trimmed K-means method[42] under the assumption that the data can be represented by well-separated spheres in variable space.

The classification method described here can be used in several ways in epidemiological studies; it could simply be seen as an exposure (eg, what happens with patients on increasing vs decreasing doses) or an outcome (eg, what interventions lead to increasing vs decreasing drug use). If the pattern classification is seen as a covariate, the investigator should weigh off the cost of adding more degrees of freedom (compared to ever-use of the drug) against the benefits of a more descriptive variable. Patients can have differential drug use based on other factors (eg, confounding by indication); causal diagram analysis during study design could identify these types of issues. This method summarizes time-varying information; if the outcome of interest occurs during the periods for which the pattern is summarized, then any detected association could be attributable to reverse causality.

There are other machine learning methods available. Auto-encoder networks are often used as automatic feature extraction methods in machine learning. Since neural networks are built on multiple layers of nonlinear calculations, the extracted features are hard to explain and often do not have a direct interpretation. As clinical interpretation of results is essential in medicine, black-box auto-encoder networks have limited use in many health-related problems. Supervised algorithms (algorithms in which patterns are optimized to distinguish between certain outcomes) could also lead to clinically relevant groups, but such classifications can then no longer be used to study the outcomes the patterns are trained on. Supervised algorithms also require a consensus about the grouping of outcomes, which may not always exist.

We illustrated this method with metformin use in diabetic patients (nondiabetics will rarely use metformin). This method can also be used for other drugs; single drug classes for chronic conditions would likely follow similar patterns and require the least amount of changes to this algorithm. In some chronic conditions, multiple drugs are indicated, for example, there are half dozen statins.[43] In some patients, a certain drug may be discontinued and another drug may be started, leading to a combined pattern of use of multiple drugs. In such cases, there are at least three options: first, one can collapse all drug classes, for example, have one statin group, and classify drug use for this combined class. Second, one can classify each drug in the class separately. Third, one can classify an expanded $N \times M$-dimensional feature space including all N features for the M different drugs in the class. Which of these approaches is most suitable will depend on the drug class under investigation, how frequently people switch between drugs in the class, and the hypothesis under investigation. This approach may need to be tweaked for drugs with more short-term use, for example, antibacterials against recurrent infections or drugs prescribed for episodic mood disorders. Such tweaks could mean including features beyond the PDC and dose variability to describe the changes in drug use. It may also not be appropriate to abstract time-varying drug use, and time-varying measures (eg, adherence measures[44]) could be included directly into a K-means clustering algorithm. If the period of use is relevant (eg, drug use guidelines changed between 1990s and 2010s), calendar time could be added as a feature as well. The principle behind the method would remain the same, regardless of the specific features used in it.

The features presented here, including F_i (the length of use), describe complex drug use which could also be used as an alternative input for complex algorithms and other machine learning applications[45,46] that use a wide variety of input data.

5 | CONCLUSION

The method to classify drug use presented in this paper allows for characterization of patient drug use into distinct and clinically relevant groups in a way that cannot be obtained from merely classifying use by quantiles of overall use.

ACKNOWLEDGMENTS

The authors acknowledge the Manitoba Centre for Health Policy for the use of data contained in the Population Health Research Data Repository under project # 2016-041 (HIPC # 2015/2016-45; REB # H2015-392; RRIC # 2015-059). The results and conclusions are those of the authors and no official endorsement by the Manitoba Centre for Health Policy, Manitoba Health, or other data providers is intended or should be inferred. Data used in this study are from the Population Health Research Data Repository housed at the Manitoba Centre for Health Policy, University of Manitoba and were derived from data provided by Manitoba Health.

DISCLOSURES

SMM has received unrestricted research grants from Merck, GlaxoSmithKline, Sanofi Pasteur, Pfizer, and Roche-Assurex for unrelated studies. SMM has received fees as an advisory board member for Sanofi Pasteur. None of the other authors has any conflict of interest to disclose.

ETHICS APPROVAL

This study was approved by the University of Manitoba Research Ethics Board and by Manitoba Health's Health Information Privacy Committee.

DATA AVAILABILITY STATEMENT

Data used in this article were derived from administrative health and social data as a secondary use. The data were provided under specific data sharing agreements only for approved use at the Manitoba Centre for Health Policy (MCHP). The original source data are not owned by the researchers or MCHP and as such

cannot be provided to a public repository. The original data source and approval for use has been noted in the acknowledgments of the article. Where necessary, source data specific to this article or project may be reviewed at MCHP with the consent of the original data providers, along with the required privacy and ethical review bodies.

ORCID

Christiaan H. Righolt https://orcid.org/0000-0002-8472-932X

REFERENCES

1. Pazzagli L, Linder M, Zhang M, et al. Methods for time-varying exposure related problems in pharmacoepidemiology: an overview. *Pharmacoepidemiol Drug Saf*. 2018;27:148-160. https://doi.org/10.1002/pds.4372
2. Strom BL, Kimmel SE, Hennessy S (eds.). Pharmacoepidemiology. Wiley-Blackwell, 2012. doi:10.1002/9781119959946.
3. Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol*. 1980;112:564-569. https://doi.org/10.1093/oxfordjournals.aje.a113025
4. Höfler M. The effect of misclassification on the estimation of association: a review. *International Journal of Methods in Psychiatric Research*. 2005;14:92-101. https://doi.org/10.1002/mpr.20
5. Brenner H, Savitz DA, Jöckel K-H, Greenland S. Effects of nondifferential exposure misclassification in ecologic studies. *Am J Epidemiol*. 1992;135:85-95. https://doi.org/10.1093/oxfordjournals.aje.a116205
6. Greenland S, Robins JM. Confounding and misclassification. *Am J Epidemiol*. 1985;122:495-506. https://doi.org/10.1093/oxfordjournals.aje.a114131
7. Vacek PM. Assessing the effect of intensity when exposure varies over time. *Stat Med*. 1997;16:505-513. 10.1002/(SICI)1097-0258(19970315)16:5<505:AID-SIM424>3.0.CO;2-Z
8. Richardson DB. Latency models for analyses of protracted exposures. *Epidemiology*. 2009;20:395-399. https://doi.org/10.1097/EDE.0b013e318194646d
9. Sylvestre M-P, Abrahamowicz M. Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Stat Med*. 2009;28:3437-3453. https://doi.org/10.1002/sim.3701
10. Berhane K, Hauptmann M, Langholz B. Using tensor product splines in modeling exposure–time–response relationships: application to the colorado plateau uranium miners cohort. *Stat Med*. 2008;27:5484-5496. https://doi.org/10.1002/sim.3354
11. Richardson DB, Terschüren C, Pohlabeln H, Jöckel K-H, Hoffmann W. Temporal patterns of association between cigarette smoking and leukemia risk. *Cancer Causes Control*. 2008;19:43-50. https://doi.org/10.1007/s10552-007-9068-7
12. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *J Roy Stat Soc: Ser C (Appl Stat)*. 1994;43:429-467. https://doi.org/10.2307/2986270
13. Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Stat Med*. 2007;26:392-408. https://doi.org/10.1002/sim.2519
14. Nagin DS. Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychol Methods*. 1999;4:139-157. https://doi.org/10.1037/1082-989X.4.2.139
15. Nagin DS, Tremblay RE. Analyzing developmental trajectories of distinct but related behaviors: a group-based method. *Psychol Methods*. 2001;6:18-34. https://doi.org/10.1037/1082-989X.6.1.18
16. Nagin DS, Jones BL, Passos VL, Tremblay RE. Group-based multi-trajectory modeling. *Stat Methods Med Res*. 2018;27:2015-2023. https://doi.org/10.1177/0962280216673085
17. Francis B, Elliott A, Weldon M. Smoothing group-based trajectory models through B-splines. *J Dev Life Course Criminology*. 2016;2:113-133. https://doi.org/10.1007/s40865-016-0025-6
18. Peristera P, Platts LG, Magnusson Hanson LL, Westerlund H. A comparison of the B-spline group-based trajectory model with the polynomial group-based trajectory model for identifying trajectories of depressive symptoms around old-age retirement. *Aging Ment Health*. 2020;24:445-452. https://doi.org/10.1080/13607863.2018.1531371
19. Lima Passos V, Crutzen R, Feder JT, Willemsen MC, Lemmens P, Hummel K. Dynamic, data-driven typologies of long-term smoking, cessation, and their correlates: Findings from the International Tobacco Control (ITC) Netherlands Survey. *Soc Sci Med*. 2019;235:112393. https://doi.org/10.1016/j.socscimed.2019.112393
20. Chu M-KM, Koval JJ. Trajectory modeling of longitudinal binary data: application of the EM algorithm for mixture models. *Commun Stat Simul Comput*. 2014;43:495-519. https://doi.org/10.1080/03610918.2012.707455
21. Nielsen JD, Rosenthal JS, Sun Y, Day DM, Bevc I, Duchesne T. Group-based criminal trajectory analysis using cross-validation criteria. *Commun Stat Theory Methods*. 2014;43:4337-4356. https://doi.org/10.1080/03610926.2012.719986
22. Goodfellow I, Bengio Y, Courville A. Deep Learning. The MIT Press, 2016.
23. Lai S, Xu L, Liu K, Zhao J. Recurrent convolutional neural networks for text classification. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI'15. Austin, Texas: AAAI Press; 2015:2267–2273.
24. Wen T-H, Gašić M, Mrkšić N, Su P-H, Vandyke D, Young S. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics; 2015:1711–1721.https://doi.org/10.18653/v1/D15-1199
25. Srivastava N, Mansimov E, Salakhutdinov R. Unsupervised learning of video representations using LSTMs. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. ICML'15. Lille, France: JMLR.org, 2015; 843–852.
26. Baytas IM, Xiao C, Zhang X, Wang F, Jain AK, Zhou J. Patient Subtyping via Time-Aware LSTM Networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '17. Halifax, NS, Canada: Association for Computing Machinery; 2017:65–74.https://doi.org/10.1145/3097983.3097997
27. Martens P, Nickel N, Lix L, et al. *The cost of smoking: a Manitoba study*. Winnipeg: Manitoba Centre for Health Policy; 2015.
28. Roos LL, Mustard CA, Nicol JP, et al. Registries and administrative data: organization and accuracy. *Med Care*. 1993;31:201-212.
29. Robinson JR, Young TK, Roos LL, Gelskey DE. Estimating the burden of disease: comparing administrative data and self-reports. *Med Care*. 1997;35:932-947.
30. Kozyrskyj AL, Mustard CA. Validation of an electronic, population-based prescription database. *Ann Pharmacother*. 1998;32:1152-1157. https://doi.org/10.1345/aph.18117
31. R: a language and environment for statistical computing. Available at: https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing. Accessed February 19, 2020.
32. WHO Collaborating Centre for Drug Statistics and Methodology. WHO. Available at: http://www.who.int/medicines/regulation/medicines-safety/about/collab-centres-norwegian/en/. Accessed February 19, 2020.
33. Steinley D. K-means clustering: A half-century synthesis. *Br J Math Stat Psychol*. 2006;59:1-34. https://doi.org/10.1348/000711005X48266

34. Selim SZ, Ismail MA. K-means-type algorithms: a generalized convergence theorem and characterization of local optimality. *IEEE Trans Pattern Anal Mach Intell*. 1984;PAMI-6:81-87. https://doi.org/10.1109/TPAMI.1984.4767478
35. Ketchen DJ, Shook CL. The application of cluster analysis in strategic management research: an analysis and critique. *Strateg Manag J*. 1996;17:441-458.
36. Bailey CJ, Turner RC. Metformin. *N Engl J Med*. 1996;334:574-579. https://doi.org/10.1056/NEJM199602293340906
37. American Diabetes Association. 9. Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes—2019. Diabetes Care 2019; 42: S90–S102. https://doi.org/10.2337/dc19-S009
38. Chen G, Khan N, Walker R, Quan H. Validating ICD coding algorithms for diabetes mellitus from administrative data. *Diabetes Res Clin Pract*. 2010;89:189-195. https://doi.org/10.1016/j.diabres.2010.03.007
39. Cobos C, Muñoz-Collazos H, Urbano-Muñoz R, Mendoza M, León E, Herrera-Viedma E. Clustering of web search results based on the cuckoo search algorithm and Balanced Bayesian Information Criterion. *Inf Sci*. 2014;281:248-264. https://doi.org/10.1016/j.ins.2014.05.047
40. Steinley D. Local optima in K-means clustering: what you don't know may hurt you. *Psychol Methods*. 2003;8:294-304. https://doi.org/10.1037/1082-989X.8.3.294
41. Hajnal I, Loosveldt G. The Effects of Initial Values and the Covariance Structure on the Recovery of some Clustering Methods. In: Kiers HAL, Rasson J-P, Groenen PJF, Schader M (eds.) Data Analysis, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization. Berlin, Heidelberg: Springer, 2000; 47–52. https://doi.org/10.1007/978-3-642-59789-3_7
42. Cuesta-Albertos JA, Gordaliza A, Matran C. Trimmed k-means: an attempt to robustify quantizers. *Ann Stat*. 1997;25:553-576.
43. Stone NJ, Robinson JG, Lichtenstein AH, et al. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults. *Circulation*. 2014;129:S1-S45. https://doi.org/10.1161/01.cir.0000437738.63853.7a
44. Steiner JF. Measuring adherence with medications: time is of the essence. *Pharmacoepidemiol Drug Saf*. 2016;25:333-335. https://doi.org/10.1002/pds.3932
45. Chubak J, Yu O, Pocobelli G, et al. Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer. *J Natl Cancer Inst*. 2012;104:931-940. https://doi.org/10.1093/jnci/djs233
46. Xu Y, Kong S, Cheung WY, et al. Development and validation of case-finding algorithms for recurrence of breast cancer using routinely collected administrative data. *BMC Cancer*. 2019;19:210. https://doi.org/10.1186/s12885-019-5432-8
47. Stewart J *Calculus*. 7th ed. Brooks/Cole, 2008.
48. Zwillinger D. CRC Standard Mathematical Tables and Formulas. CRC Press. Available at: https://www.crcpress.com/CRC-Standard-Mathematical-Tables-and-Formulas/Zwillinger/p/book/9781498777803. Accessed February 19, 2020

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Righolt CH, Zhang G, Mahmud SM. Classification of drug use patterns. *Pharmacol Res Perspect*. 2020;e00687. https://doi.org/10.1002/prp2.687

APPENDIX A

DERIVATION OF FEATURES

A.1 | DOSE CHANGE

The overall trend of use can be described by a linear fit of use over time (using only the start and end dose would not describe the pattern in between). Without loss of generality, we can write the relation between the dose and exposure episode as:

$$d_{ij} = \alpha_i + \beta_i p_{ij} + \varepsilon_{ij}, \tag{6}$$

where α_i and β_i are the person-specific intercept and slope and ε_{ij} captures both nonlinear elements and random fluctuations. When we use $\vec{p}_i = (1, 2, 3, \ldots, N_i)$ for the episodes and $\vec{d}_i = (d_1, d_2, d_3, \ldots, d_{N_i})$ for the doses in those episodes, the least squares estimate for β_i is:

$$\beta_i = \frac{\mathrm{cov}(\vec{p}_i, \vec{d}_i)}{\mathrm{var}(\vec{p}_i)}, \tag{7}$$

where *var* and *cov* denote variance and covariance, respectively. The slope itself is not normalized to the numbers of episodes (which causes differences based on τ), and we define the trend, T_i, as the normalized measure

$$T_i = N_i \beta_i \tag{8}$$

We can derive a closed form formula for $\mathrm{var}(\vec{p}_i)$, when we note that

$$\sum_{n=1}^{N} n = \frac{N(N+1)}{2} \tag{9}$$

And

$$\sum_{n=1}^{N} n^2 = \frac{2N^3 + 3N^2 + N}{6}, \tag{10}$$

see for example pages A35-A36 in.[47] When we define the mean of all p_{ij} for a person as , we can derive

$$\mathrm{var}(\vec{p}_i) = \frac{1}{N_j} \sum_{j=1}^{N_i} [p_{ij} - \vec{p}_i]^2$$

$$= \frac{1}{N_i} \sum_{n=1}^{N_i} \left[n - \frac{1}{N_i} \sum_{m=1}^{N_i} m \right]^2 \tag{11}$$

$$= \frac{1}{N_i} \sum_{n=1}^{N_i} n^2 - \frac{(N_i + 1)^2}{4}$$

$$= \frac{N_i^2 - 1}{12}$$

When we substitute these results in Equations 7 and 8, we get

$$T_i = \frac{12N_i}{N_i^2 - 1}\mathrm{cov}(\bar{\bar{d}}_i, \bar{p}_i) \tag{12}$$

Note that when all persons have the same follow-up period (N_i=N for all i), the term $\frac{12N_i}{N_i^2-1}$ is constant and $\mathrm{cov}(\bar{\bar{d}}_i, \bar{p}_i)$ can directly be used as a feature for the dose change as well.

One could argue that the slope β_i could be a measure of change or trend itself as well. But this slope is dependent on τ and does not allow for direct comparison between analyses in which τ varies. For example, β_i would become one-third of the value when 30-day periods are examined instead of 90-day periods, but T_i remains the same (one-third of the period length leads to three times as many periods, which leads to one-third of the slope, before normalization by the number of periods). Running the clustering algorithm is computational-intensive and requires a large generalizable patient population. Doing this kind of analysis may not always be feasible, for example, for prospectively collected data of small to medium sized cohorts. If clustering results are available for drug use in a general population, patients in a secondary analysis could still be assigned to their respective patient group, even if the length of reporting periods of drug use do not equal those periods of the available clusters.

A.2 | DOSE VARIABILITY

The variability in the prescribed strength is a measure of stability of use. The dose change describes the overall change, and the variability V_i here describes the fluctuations in dose (ie, a discrete version of the second derivative),

$$V_i = \frac{1}{N_i - 2}\sum_{j=2}^{N_i-1}\left|2\frac{d_{ij}}{t_{ij}} - \frac{d_{i(j-1)}}{t_{i(j-1)}} - \frac{d_{i(j+1)}}{t_{i(j+1)}}\right| \tag{13}$$

We use the average absolute value of the discrete second derivative $2\frac{d_{ij}}{t_{ij}} - \frac{d_{i(j-1)}}{t_{i(j-1)}} - \frac{d_{i(j+1)}}{t_{i(j+1)}}$ (the discrete version, without error term, of the expression on page 764 in [48]). We use a factor $N_i - 2$, because the second derivative is not defined for the first and last episodes.

One could argue that the variance of a dose sequence characterizes the variability. The difference between equation 13 and the variance (of d_{ij} for person i) is that equation 13 sums the absolute value of the second-order difference, while variance reflects the spread around the mean value. For an increasing dose (a straight line with nonzero slope), V_i=0 (because there is no point where the slope changes sign), but the variance is large (characterizing the trend rather than the variability). The variability measure V_i is intended to capture this second-order variation on top of the dose change T_i.