Markov-modulated marked Poisson processes for modeling disease dynamics based on medical claims data

Sina Mews1 | Bastian Surmann2 | Lena Hasemann2 | Svenja Elkenkamp2

1Department of Business Administration and Economics, Bielefeld University, Bielefeld, Germany
2Department for Health Economics and Health Care Management, Bielefeld University, Bielefeld, Germany

Correspondence
Sina Mews, Department of Business Administration and Economics, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany.
Email: sina.mews@uni-bielefeld.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: Projektnummer 316099922-TRR 212

We explore Markov-modulated marked Poisson processes (MMMPs) as a natural framework for modeling patients’ disease dynamics over time based on medical claims data. In claims data, observations do not only occur at random points in time but are also informative, that is, driven by unobserved disease levels, as poor health conditions usually lead to more frequent interactions with the health care system. Therefore, we model the observation process as a Markov-modulated Poisson process, where the rate of health care interactions is governed by a continuous-time Markov chain. Its states serve as proxies for the patients’ latent disease levels and further determine the distribution of additional data collected at each observation time, the so-called marks. Overall, MMMPs jointly model observations and their informative time points by comprising two state-dependent processes: the observation process (corresponding to the event times) and the mark process (corresponding to event-specific information), which both depend on the underlying states. The approach is illustrated using claims data from patients diagnosed with chronic obstructive pulmonary disease by modeling their drug use and the interval lengths between consecutive physician consultations. The results indicate that MMMPs are able to detect distinct patterns of health care utilization related to disease processes and reveal interindividual differences in the state-switching dynamics.

KEYWORDS
chronic obstructive pulmonary disease, continuous time, disease process, hidden Markov model, informative observation times, maximum likelihood

1 INTRODUCTION

Whenever a patient interacts with the health care system, claims data are routinely collected from all providers caring for the patient (such as physicians, pharmacies, or hospitals), containing information on costs, patient-specific diagnoses, and received treatments like medication. These rich databases on real-life health care provisions are increasingly prominent in public health research as well as decision-making processes of different stakeholders. For example, claims data are used to optimize health service provisions, estimate the prevalence and incidence of diseases, or identify patients at risk of hospital readmission. While claims data have been extensively used for disease prediction, much less attention has been paid to their vast amount of (implicit) information on disease dynamics over time, such as (changes in) medication,
hospital stays, or the frequency of physician consultations (but see Ploner et al\textsuperscript{17}). In this contribution, we thus draw on these comprehensive datasets to model the temporal courses of diseases and to learn about patients’ health conditions over time. Gaining insights into the latter is crucial to assess treatment effects on disease progression and to distinguish health conditions associated with different demands of care. These results, in turn, can be used to evaluate individual as well as economic consequences of specific diseases.

For modeling patients’ disease processes over time, medical claims data pose two main challenges, namely (1) that patients’ disease stages are not (directly) observed and (2) that patients’ interactions with the health care system are driven by the disease stages themselves, with more frequent interactions when the patient’s condition is poor. Regarding (1), while claims data provide extensive records on patients’ received services and diagnoses, they lack information on the actual health condition, specifically on distinct disease stages (such as mild, moderate, and severe) graded according to disease-specific practice guidelines. Therefore, the unobserved evolution of patients’ disease activity over time needs to be inferred from available medical data. A popular approach for estimating properties of the disease process underlying the observed data are (continuous-time) hidden Markov models (HMMs).\textsuperscript{8,9} Although these latent-state models have been successfully applied in different studies on disease progression, for example based on monthly MRI scans\textsuperscript{10} or screening data,\textsuperscript{11} they are not appropriate for modeling routinely collected claims data due to the second (2) challenge: whereas the occurrence of earthquakes,\textsuperscript{14} if additional event-specific information, so-called marks, are collected at each event time, then both can be modeled jointly using Markov-modulated Poisson processes (MMPPs). MMPPs generalize homogeneous Poisson point processes in that they assume the event rate to depend on a latent finite-state Markov process in continuous time. According to this underlying discrete-valued Markov process, the event rates switch between different levels over time, thus resulting in clustering of events. These events can, for example, correspond to mouse clicks (i.e., Web page requests),\textsuperscript{12} the surfacing of whales,\textsuperscript{13} or the occurrence of earthquakes.\textsuperscript{14} If additional event-specific information, so-called marks, are collected at each event time, then both can be modeled jointly using Markov-modulated marked Poisson processes (MMPMs). These models thus consist of three stochastic processes: the state process, the observation process, and the mark process. The state process, corresponding to the unobserved Markov process, governs both the observation process, consisting of the observation times, and the mark process, consisting of the data collected at each observation time (i.e., the marks). In particular, the distributions of the marks and of the time intervals between consecutive observations are both determined by which latent, discrete-valued state is active at the time.

While MMPPs have been applied in various areas—for example, to model occurrence patterns of earthquakes,\textsuperscript{14} opportunistic capture-recapture data\textsuperscript{15} as well as check-in\textsuperscript{16} and clickstream data\textsuperscript{17}—applications in the medical context and especially in disease modeling are rare. Lange et al\textsuperscript{18} introduced the use of MMPPs to jointly model patients’ visit times and their observed (though possibly misclassified) disease status at each observation time based on electronic health records (EHR) data. Their motivating example was screening for mammography in women with a first diagnosis of breast cancer, and the observed data consisted of the visit times and the mammography result at each time, which they modeled as an MMMPP. Furthermore, Alaa et al\textsuperscript{19} used a similar modeling approach in the medical context, namely the semi-Markov-modulated marked Hawkes process, to infer hospitalized patients’ latent clinical states over time. In particular, they modeled a sequence of informatively sampled physiological data from critically ill patients in regular wards to assess their risk of clinical deterioration and prevent delayed transfer to intensive care units.

This contribution presents claims data as another key application for MMPPs, since it is reasonable to assume that increased disease activity results in higher health care utilization.\textsuperscript{18–21} Consequently, claims data contain information on the underlying disease process by accurately reflecting (the frequency of) patients’ interactions with the health care system, which often occur in clusters (see Figure 1). These health care interactions can thus be modeled as an MMPP, where the rate of interactions is governed by a continuous-time Markov chain, whose states serve as a proxy for the patients’ unobserved disease levels. At each interaction time with the health care system, claims data provide additional observations (i.e., marks), such as the amount of drug use or costs associated with each health care interaction, that can help to improve inference on the underlying disease process. In particular, by allowing different observations to be considered in the mark process, we extend the method of Lange et al\textsuperscript{18} to a more general representation of the marks’ state-dependent distributions. Regarding Alaa et al,\textsuperscript{19} the main difference to our approach is that they focus on risk prognosis based on EHR data, particularly patients’ vital signs and lab tests, whereas we aim to extract information on patients’ health conditions over time from claims data. While claims data pose particular challenges as outlined above, these can be addressed...
naturally by MMMPPs, which allow inference on the latent disease process based on observations occurring at informative and clustered points in time.

2. **MARKOV-MODULATED MARKED POISSON PROCESSES**

2.1. **Basic model formulation**

We consider (claims) data containing information on the random observation times \( T_0, T_1, \ldots, T_n \), \( 0 = T_0 < T_1 < \ldots < T_n \), which occur at irregularly spaced points in time, as well as additional data \( Y_{t_0}, \ldots, Y_{t_n} \) collected at the realized observation times. These sequences of random variables are referred to as the observation process and the mark process, respectively, and depend on an underlying, unobserved state process \( \{ S_t \}_{t \geq 0} \). From now on, let the integer \( \tau = 0, 1, \ldots, n \) denote the index of the observation in the sequence, such that \( Y_{\tau} \) and \( S_{\tau} \) shorten to \( Y_\tau \) and \( S_\tau \), respectively.

The state process is modeled as an \( N \)-state continuous-time Markov chain. Transitions between the states are governed by a transition intensity matrix \( Q = (q_{ij})_{i,j=1,\ldots,N} \), whose off-diagonal elements \( q_{ij} \geq 0, i,j = 1, \ldots, N, i \neq j \), can be interpreted as the rates at which transitions from state \( i \) to state \( j \) occur. The duration in each state \( i = 1, \ldots, N \) is exponentially distributed with parameter \( \lambda_i \), where \( \lambda_i \) is the \( i \)th diagonal entry in \( Q \). Furthermore, the initial distribution of the state process is denoted by \( \delta = (\delta_1, \ldots, \delta_N) \), where \( \delta_i = \Pr(S_0 = i) \). In the context of claims data, the states represent disease levels associated with different frequencies of health care interactions (as reflected in the observation process) and different distributions of variables collected at each observation time (as reflected in the mark process). It is important to note that although the states can approximate distinct disease levels, they are purely data-driven, as the model just picks up the strongest patterns in the data. With no information on patients’ actual health condition available, the states thus do not (necessarily) correspond to any formally defined disease stages based on practice guidelines or grading systems.

The observation process is modeled as a doubly stochastic point process, namely an MMPP whose event rates \( \lambda_i, i = 1, \ldots, N \), are selected by the underlying Markov chain. Conditional on the current state \( i = 1, \ldots, N \) being active throughout the time interval \([0, t]\), the number of events in \([0, t]\) follows a Poisson distribution with parameter \( \lambda_i t \). Furthermore, the waiting times between consecutive events \( X_\tau = T_\tau - T_{\tau-1}, \tau = 1, \ldots, n \), are exponentially distributed with parameter \( \lambda_i \) within each state. In the context of modeling medical claims, the frequency, or rate, of observations (ie, health care interactions) \( \lambda_i \) is thus determined by the underlying (disease) state \( i = 1, \ldots, N \). We assume that severe disease levels are associated with higher event rates \( \lambda_i \), as an increased disease activity (usually) causes higher health

---

**Figure 1** Example daily defined dose sequence, that is, a measure of drug use, (upper plot) and step function of the number of consultations over time (lower plot) for one patient.
likelihood evaluation and maximization

For the mark process, we assume the distribution of a variable (ie, mark) \( Y_t \) collected at observation time \( t \) to be fully determined by its underlying state. In particular, the conditional independence assumption \( f(y_t | t_1, \ldots, t_r, y_1, \ldots, y_{r-1}, s_1, \ldots, s_r) = f(y_t | s_r) \) implies that the underlying Markov chain selects which state-dependent distribution \( f(y_t | s_r = i) \) is active at time \( t_r \). As the state-dependent distributions can take on any (parametric) form, various data types, such as binary, count, or continuous variables, can be considered in the mark process. Note that while the marks are conditionally independent of all previous (and future) marks, observation times, and states, the state process induces correlation in the mark (as well as the observation) process. For now, we assume all stochastic processes to be homogeneous, but will discuss how to relax this assumption in Section 2.3.

### 2.2 Likelihood evaluation and maximization

Subject to the unobserved Markovian state process, the MMMPP jointly models the mark process and the observation process. To evaluate the corresponding likelihood of the model, inferential tools from the HMM framework, in particular the corresponding efficient algorithms for parameter estimation, can be applied. In contrast to Lu et al, who estimate the model parameters based on the EM algorithm, we numerically maximize the likelihood using the HMM-based forward algorithm. Defining the observation process by its waiting times, the likelihood of the observed sequence \( \{(x_t, y_t)\}_{t \in \{0,1,\ldots,n\}} \) is given by:

\[
\mathcal{L} = \delta \mathbf{P}(y_0) \left( \prod_{r=1}^{n} \exp((Q - \Lambda)x_r) \Lambda \mathbf{P}(y_r) \right) \mathbf{1},
\]

where \( \mathbf{P}(y_r) = \text{diag} \{ \Pr(y_r | s_r = 1), \ldots, \Pr(y_r | s_r = N) \} \) and \( \Lambda = \text{diag} \{ \lambda_1, \ldots, \lambda_N \} \) are diagonal matrices and \( \mathbf{1} \in \mathbb{R}^N \) denotes a column vector of ones. Note that different methods to compute the matrix exponential \( \exp(\mathbf{A}) = \sum_{d=0}^{\infty} \mathbf{A}^d/d! \) can be used. Equation (1) corresponds to a recursive calculation of the likelihood, where the joint likelihood of the observed sequence up to time \( t_r \) is updated based on the likelihood up to time \( t_{r-1} \), retaining information on the probabilities of the different states being active. To avoid numerical under- or overflow in the maximization, a scaling strategy needs to be applied (see Zucchini et al for further details on technical issues arising in the likelihood maximization).

### 2.3 Incorporating covariates

Covariates can be incorporated in all three components of the MMMPP using a general linear regression framework. Depending on whether the covariates affect the switching dynamics of the state process, the event frequency of the observation process, or the state-dependent distributions of the mark process, the respective parameters can be modeled as a function of both individual-specific and time-varying covariates.

Regarding the mark process, for example the mean of the state-dependent distributions can be specified as \( \mu_{ij}^{(t)} = g^{-1}(\beta_0^{(t)} + \beta_1^{(t)} z_t) \), where \( z_t \) is a time-varying covariate, \( \beta_0^{(t)} \) and \( \beta_1^{(t)} \) are state-dependent coefficients, and \( g^{-1}(\cdot) \) denotes an inverse link function accounting for possible parameter constraints. In contrast to the mark process, modeling the state process or the observation process as a function of covariates is straightforward only if covariate values are constant across time. For instance, given a time-constant, individual-specific covariate \( z_{(k)} \) for the \( k \)-th person in the dataset, the (off-diagonal) state transition intensities can be specified as \( q_{ij}^{(k)} = \exp(\beta_0^{(t)} + \beta_1^{(t)} z_{(k)}), \) and the event rates as \( \lambda_{ij}^{(k)} = \exp(\beta_0^{(t)} + \beta_1^{(t)} z_{(k)}) \), respectively. In this case, the parameters of the respective process differ between individuals, but importantly, they remain constant over time. If instead, a covariate varies (continuously) across time, leading to (either) nonhomogeneous state transition intensities or event rates, the resulting matrix exponentials in Equation (1) are not analytically tractable. However, the latter can be calculated recursively over shorter intervals with constant parameter values. For time-varying covariates that are piecewise constant over time—such as the age of a person measured in years—the matrix exponentials are thus calculated separately for each time interval on which the covariate, and hence the parameters \( q_{ij}^{(t)} \) or \( \lambda_{ij}^{(t)} \), respectively, are constant.
diagnosis—their effects and hence the respective parameters \( q_{ij}^{(t)} \) or \( \lambda_i^{(t)} \) can be approximated by constant step functions (see, eg, Langrock et al\(^{13}\) or Mews et al\(^{25}\) for a detailed account of the approach).

### 3 | CASE STUDY: MODELING DISEASE ACTIVITY OF COPD PATIENTS

#### 3.1 | Data description and model formulation

We consider data from one of the largest statutory health insurance (SHI) companies in Germany, providing comprehensive information on interactions of the insured with the health care system. Covering a period of 16 years, namely from 2005 to 2020, the data comprise, inter alia, the insured persons’ sex and age as well as their diagnoses (based on ICD-10 classification), prescriptions, hospitalizations, sick days, and associated health care costs. Consequently, the data set allows not only to identify particular groups of patients suffering from a certain disease but also to derive details on their health condition. In this case study, we focus on patients diagnosed with chronic obstructive pulmonary disease (COPD), which was previously analyzed in similar studies on latent-state modeling.\(^{26}\) COPD is a common lung disease with a prevalence of 6.37% in Germany in 2017.\(^{27}\) It is associated with persistent respiratory symptoms and can be treated to slow, though not fully reverse its progression. Due to its (usually) slow progression, longitudinal data are necessary to gain insight into the disease process of COPD over time.

To cover the longest possible observation period and to reduce heterogeneity in our study population, we consider persons initially diagnosed with COPD in 2008 (using the years 2005-2007 as a pre-observation period; see Section A in Appendix S1 for detailed information on the criteria used to select the study population). As COPD is often associated with various comorbidities, which hinder inference on the disease dynamics, we calculate the age-adjusted Charlson comorbidity index (ACCI) based on Quan et al\(^{28}\) for each person. Data from persons with an ACCI maximum score larger than four (corresponding to severe age-adjusted comorbidities)\(^{29}\) were excluded from analyses. As a consequence, our study population is younger than COPD patients in the general population. The final dataset includes 470 persons (141 males and 329 females) with a median age of 49 years (min: 21; max: 69) at initial diagnosis. Regarding the ACCI, most persons (80%) show a moderate age-adjusted comorbidity burden (ie, an ACCI of 3 or 4), while the rest (20%) have mild age-adjusted comorbidities (ie, an ACCI of 1 or 2).\(^{29}\) Overall, the final data comprise 112,297 observations, covering a mean observation period of 12 years per person (min: 6; max: 13) after initial COPD diagnosis.

For modeling COPD patients’ disease activity over time, we consider both the interval length between consecutive physician consultations (ie, the waiting times) and patients’ drug use. The latter is measured in daily defined doses (DDDs)—a standardized unit for drug consumption—based on physicians’ prescriptions contained in the SHI data. As not every consultation is associated with a drug prescription, we have a large amount of zero DDDs (ie, no prescribed drugs; 59.5%) in our data. If the patient received a prescription, the average amount of DDDs prescribed is 106 (SD: 115). Regarding the waiting times, on average 18 days pass between consecutive physician consultations (SD: 25). Histograms indicating the distribution of the observed variables are shown in Figure S1 in Appendix S1. Furthermore, an example trajectory of a patient’s DDDs over time together with the accumulative number of consultations is shown in Figure 1, reflecting the clustered structure of the observations. In particular, one can easily distinguish periods with high (eg, 12 consultations within 31 days) and low health care utilization (eg, more than 1 year between consecutive consultations). The relation between the frequency of physician consultations and DDDs, in contrast, is not as obvious: there is a tendency to observe more consultations without drug prescription (ie, zero DDDs) if the frequency of interactions is high, but no clear patterns regarding the amount of DDDs prescribed is evident.

#### 3.2 | Model formulation

We model the observed waiting times between consecutive physician consultations and the DDDs prescribed as an MMMPP, where the underlying states can be interpreted as patients’ health condition over time. Regarding the observation process, this means the waiting times are modeled using state-dependent exponential distributions, while for the mark process, we assume the DDDs to follow a zero-adjusted gamma distribution with state-specific parameters. As we are interested in interindividual differences in the state-switching dynamics, we model the state transition intensities as
a function of patients' sex, their age at initial diagnosis (ageD), and the ACCI (dichotomized into either mild or moderate age-adjusted comorbidities):

\[ q_{ij} = \exp\left( \beta^{(ij)}_0 + \beta^{(ij)}_1 \text{sex} + \beta^{(ij)}_2 \text{ACCI} + \beta^{(ij)}_3 \text{ageD} \right), \quad \text{for} \quad i \neq j. \]

For simplicity, we restrict ourselves to a two-state model, noting that the methodology is generally applicable for any finite number of states. Arguably, this is only one of many possible model formulations that can be considered (cf. Section 4)—however, presenting a comprehensive analysis of COPD patients' disease progression is beyond the scope of this case study. The parameters of the MMMP, namely the regression coefficients and the state-dependent rate parameters as well as the parameters of the gamma distributions, are estimated by numerically maximizing the joint likelihood over all patients, which is calculated as the product of the individual likelihoods given in Equation (1). The R code for model fitting, subsequent state decoding, and data simulation, as well as to reproduce the tables and figures presented in this article is available in Appendix S2.

### 3.3 Results

The parameter estimates of both the observation and the mark process are presented in Table 1, while plots visualizing the estimated state-dependent distributions are shown in Figure S1 in Appendix S1. The small 95% CIs of the estimated parameters reflect that uncertainty within the estimated densities is very low, which is due to the large sample size. Regarding the observation process, the rate of health care interactions in state 1 is roughly $1/8$, whereas in state 2, it is $1/34$. Therefore, we would expect one consultation every 8 days in state 1 compared to one consultation every 34 days in state 2. Regarding the mark process, while the estimated gamma distribution in state 1 comprises, on average, lower DDDs than the one in state 2, the estimated variances in both states are high. Consequently, there is substantial overlap of both estimated gamma distributions. In state 1, however, the probability of observing no prescription (ie, a zero DDD) is much higher than in state 2, namely by 36 percentage points. The latter appears reasonable: if a patient has many physician consultations in a short time interval (eg, because of different examinations or tests), they will not receive a prescription at each of these visits. Taking into account the state-dependent interaction rates, the expected sum of DDDs per month (30 days) is 245 in state 1 compared to 208 in state 2, indicating that although the amount of prescribed drugs at a single consultation is expected to be lower in state 1, the higher interaction rate leads to a larger expected number of DDDs than in state 2. Overall, state 1 is thus characterized by frequent health care interactions and higher drug use (over a longer time period), which could be interpreted as a state of poor health condition or, alternatively, a period in which a treatment needs to be appropriately adjusted to a patient (eg, right after disease diagnosis)—an interpretation supported by the estimated initial state probabilities, as it is 2.5 times more likely that a person is in state 1 rather than state 2 right after their initial COPD diagnosis. In contrast, state 2 consists, on average, of approximately monthly health care interactions with higher DDDs at a single consultation and hence could be interpreted as a state in which a patient’s therapy is maintained. Therefore, we (tentatively) describe state 1 as the high and state 2 as the low disease level. Importantly, however, these disease states are derived in a data-driven way and as such should not be expected to match disease stages postulated in the literature exactly.

### Table 1

(State-dependent) parameter estimates with 95% confidence intervals (CIs) for the initial state distribution, the observation process, and the mark process, where the (zero-adjusted) gamma distribution is parameterised in terms of its mean and SD.

| Parameter                  | Estimate       | 95% CI       |
|----------------------------|----------------|--------------|
|                            | State 1        | State 2      | State 1        | State 2      |
| Initial state prob. $\delta_i$ | 0.707          | 0.293        | [0.633; 0.771] | [0.229; 0.367] |
| Rate of exponential dist. $\lambda_i$ | 0.123          | 0.029        | [0.121; 0.125] | [0.029; 0.030] |
| Mean of gamma dist. $\mu_i$ | 80.9           | 124.5        | [79.1; 82.7]   | [122.8; 126.3] |
| SD of gamma dist. $\sigma_i$ | 85.0           | 116.9        | [82.8; 87.2]   | [115.0; 118.8] |
| Prob. at zero $\pi_0^{(i)}$ | 0.726          | 0.370        | [0.721; 0.730] | [0.362; 0.379] |

Note: The CIs were calculated based on the observed Fisher information.
TABLE 2 Expected duration (in days) as well as expected percentage of time spent in each state with 95% CIs, calculated for different groups (based on the estimated regression coefficients presented in Table S1 in Appendix S1).

| Group          | Expected duration (in days) | Expected percentage of time |
|----------------|----------------------------|----------------------------|
|                | State 1 | State 2 | State 1 | State 2 |
| Reference group| 48.2    | [45;51.7] | 94.2    | [88.3;100.3] | 33.8%    | [32.8;35.0] | 66.2%    | [65.0;67.2] |
| Male           | 49.1    | [44.7;54.3] | 175.1   | [158.4;194.4] | 21.9%    | [20.7;23.2] | 78.1%    | [76.8;79.3] |
| Mild ACCI      | 53.5    | [47.3;60.6] | 148.3   | [131.7;167.8] | 26.5%    | [24.9;28.3] | 73.5%    | [71.7;75.1] |
| Min. age: 21   | 61.3    | [52.0;72.7] | 77.0    | [65.9;90.1] | 44.3%    | [41.5;47.2] | 55.7%    | [52.8;58.5] |
| Max. age: 69   | 40.7    | [36.2;45.7] | 108.4   | [96.6;121.4] | 27.3%    | [25.7;28.8] | 72.7%    | [71.2;74.3] |

Notes: CIs were obtained based on Monte Carlo simulation from the estimators’ approximate distribution as implied by maximum likelihood theory. The reference group are female patients with moderate comorbidities and age 49 at initial diagnosis (corresponding to the mean age at diagnosis in the study population). For all other groups, only the specified characteristic is changed with regard to the reference group (for instance, the second row are male patients with moderate ACCI and mean age at diagnosis).

The estimated regression coefficients governing the state process are presented in Table S1 in Appendix S1. From these, we can calculate the transition intensity matrix $Q$ for different covariate values, from which in turn we can derive the expected durations within each state (cf. Section 2.1). Results are presented in Table 2. The reference group of female patients with moderate comorbidities and age 49 at initial diagnosis (corresponding to the mean age at diagnosis in the study population) is expected to spend roughly 48 days in state 1 compared to 94 days in state 2. Thus, we would expect a person from the reference group to spend approximately one third of her time in the high disease level and two thirds in the low disease level (corresponding to the stationary distribution of the latent Markov chain, given the respective covariate values). In contrast, male patients with otherwise the same characteristics spend considerably more time in the low disease level (ie, 11.9 percentage points more), mainly because they remain in state 2 for a longer time period before switching to state 1. A similar pattern is found for mild (in contrast to moderate) comorbidities, with a difference of 7.3 percentage points in the expected time spent in each state. Regarding age at diagnosis, there is a clear effect that younger persons spend both more time in state 1 and less time in state 2 compared to the reference group, which is reversed in older age (cf. Figure S2 in Appendix S1, which plots the expected percentage of time spent in the high disease level for the whole age range observed in the study population).

Based on the fitted MMMPP, we can use the Viterbi algorithm to infer the most probable sequence of latent (disease) states for each person. These decoded state sequences provide insight into the individual course of a disease by classifying each observation into either the high or low disease level (see Figure 2). In addition, we calculated the local state probabilities using the forward-backward algorithm, which allow us to quantify uncertainty in the decoded state sequence (cf. Figure S3 in Appendix S1). Overall, most observations in the SHI data (namely 69.3%) are assigned to one of the two states with a probability higher than 80%. This proportion is higher for state 1 compared to state 2 (cf. Figure S4 in Appendix S1). Given that the state-dependent distributions of the observation process and the mark process overlap substantially, uncertainty in state allocation is to be expected and moreover, reflects that disease dynamics within the data are less distinct. Nevertheless, the model appears to adequately distinguish (qualitatively) different periods of health care utilization and drug use, while taking into account the temporal dependence structure of the data.

To further assess the model fit, we additionally fitted a simple (ie, single-state) marked Poisson process to the data and then simulated observations based on both estimated models (ie, the simple model and the MMMPP) to check whether they could reproduce the patterns found in the SHI data. The resulting synthetic data of both models adequately reflect the distribution of the DDDs, indicating that a single state seems sufficient to describe the marks (cf. Figures S5 and S6 in Appendix S1). Regarding the observation process, however, (consecutive) short and long waiting times are underrepresented when simulating from the simple model, whereas simulations from the MMMPP match the distribution of the real waiting times well. The more flexible Markov-modulated modeling approach is thus needed to adequately capture the patterns of the SHI data and especially periods with varying frequencies of health care interactions, which is also reflected in the BIC values of both models (simple model: 1 533 714; MMMPP: 1 506 409). Furthermore, comparing simulated state sequences from the MMMPP to the decoded ones of the SHI data shows similar dynamics over time, implying that the MMMPP can reproduce reasonable (disease) state trajectories for COPD patients. Regarding the tails of the distributions of simulated DDDs and waiting times, however, the MMMPP as well as the simple model generate less
extreme values than can be observed in the real data, hence indicating that more heavy-tailed distributions might improve the model fit.

### 3.4 Limitations and discussion of findings

When interpreting the model results, it is important to keep in mind that the physician consultations and DDDs considered in the analysis are not exclusively related to COPD, but comprise all health care interactions of a patient. Consequently, the (disease) states identified correspond to general health conditions of COPD patients, meaning that a higher disease level (i.e., state 1) does not necessarily correspond to a worsening of COPD but could relate to any (temporary) poor health condition. This allows us to capture a wide range of COPD-related health issues. To be able to draw any medical conclusions or practical implications from the case study, however, the data basis would need to be refined and the model further developed as well as validated.

As the state-dependent distributions of the mark process overlap substantially (except for the probabilities of zeros), state decoding is mainly driven by the waiting times and less by the DDDs. It is thus unclear if the states merely distinguish periods of high and low health care interactions or if they provide further information on the disease process by accounting for drug use. In particular, the informative value of DDDs for the disease severity is questionable, given that some vitamin supplements can have much higher DDDs than COPD-specific drugs. However, information on specific clinical measurements like biomarkers is unavailable, since claims data are collected for billing purposes and not originally intended for scientific research. Yet, other variables contained in claims data—such as costs associated with each health care interaction or specific treatments—could (additional) be used in the mark process to facilitate inference on the underlying disease levels. Furthermore, to validate the interpretation of the states as (distinct) disease levels, the decoded state sequences of the fitted MMMPP would need to be compared to additional data on persons’ true health status over time.

Regarding the estimated covariate effects in our model, the result that persons with mild ACCI spend less time in state 1 compared to those with moderate ACCI reinforces the interpretation of state 1 as a higher disease level than state 2. Of all considered groups, male patients are expected to remain in state 2 longest, which is characterized by less frequent health care interactions, before switching to state 1—whether this is caused by different underlying disease dynamics or by the fact that men tend to wait longer before seeking medical advice than women cannot be answered by our model. Furthermore, it may appear puzzling that the expected percentage of time in the high disease state
considerably decreases as a function of age at diagnosis (while accounting for patients’ sex and ACCI). This effect, however, is likely confounded by our use of the age-adjusted comorbidity index (ACCI): since the ACCI adds one additional point for each decade after fifty, persons in their sixties, for instance, are only allowed two additional comorbidities to remain in the study population (as we excluded everyone with an ACCI larger than 4), whereas persons younger than 50 can have up to four additional comorbidities. As persons with more comorbidities have a higher health care utilization, the age-specific difference in (known) comorbidities thus leads to younger patients spending more time in state 1.

4 | DISCUSSION

When working with longitudinal data to understand the evolution of disease processes, a high-level decision to be made is whether observation times are (assumed to be) noninformative or informative (ie, dependent on the measure of interest, such as the disease severity). In particular, neglecting an informative observation process in the analysis of disease dynamics or outcomes potentially leads to biased parameter estimates. This risk of bias, however, appears to be lacking in awareness, since health care longitudinal studies rarely report on the potential informativeness of observation times. While joint models incorporating both informative observation times and disease processes exist, they mostly rely on data with directly observed disease stages and assume prescribed examinations with informative missingness instead of patient-initiated visit times. Specifically for EHRs and medical claims data, however, observation times are likely correlated with patients’ latent disease dynamics: for example, patients with more severe conditions or acute symptoms possibly visit their physicians more often than those with mild conditions or no symptoms. In addition to the observation times, observed variables like patients’ drug consumption or health care costs depend on the disease status, which is unobserved in claims data. Therefore, we propose to infer disease dynamics from the latter using Markov-modulated marked Poisson processes. Instead of regarding the informative observation process as a nuisance that needs to be accounted for, MMMPPs explicitly use it to infer latent (disease) states by integrating both the event times (ie, the observation times of health care interactions) and event-specific information (ie, the observed marks) into a joint model.

Due to their flexible model structure, MMMPPs offer manifold possibilities to analyze claims data; they not only operate in continuous time and allow for (potentially multivariate) observations consisting of various data types but can also include covariate effects on the disease dynamics, the event rates, or the mark distributions. In many applications, however, it is not a priori clear which parameters, that is, which model components, depend on covariates. For example, persons’ sex might directly affect their event rate—for example, women might have more frequent interactions with the health care system than men—or rather the disease dynamics, which then result in varying event rates—for example, women might spend more time in poor health states than men, resulting in more frequent health care interactions. While model selection methods can be applied to choose between different candidate models, this task already becomes impractical for a moderate number of covariates, given that each can possibly affect all three processes. In addition, model selection is not only a challenge regarding covariates but also regarding the number of (disease) states underlying the observation sequences. As medical claims are complex data sets characterized by, for example, outliers, multimodality, and individual heterogeneity, additional states can capture otherwise neglected structure in the data and thus, significantly improve the model likelihood, which is why information criteria such as AIC and BIC usually point to models with undesirably high numbers of states—a problem well known for HMMs. Therefore, we recommend to be pragmatic about both including covariates and selecting the number of (disease) states by taking into account expert knowledge, the current state of research, computational considerations, and especially the specific research question(s) tackled. To include a covariate in only one of the possible processes and to select a small, reasonable number of distinct states can help retain the model’s interpretability.

Although the MMMPPs presented incorporate significant characteristics of claims data, such as the informative observation process, they do not yet allow for unobserved individual heterogeneity or a flexible distribution of the duration in the states. While both issues are particularly relevant in the medical context, including random effects and/or (continuous-time) semi-Markov processes in the model proves challenging. For numerically maximizing the likelihood, integrating over all possible values of random effects renders the likelihood calculation intractable, which is why other estimation techniques such as Bayesian methods or the Monte-Carlo EM algorithm are required. Furthermore, (continuous-time) semi-Markov processes are generally more demanding to apply than models assuming the Markov
property,\textsuperscript{19,46,47} while incorporating covariates is not straightforward anymore.\textsuperscript{48} Despite this potential for model extensions, MMMPPs are able to detect distinct patterns of health care utilization related to disease dynamics and their (possible) association with person characteristics, as illustrated in the case study on statutory health insurance data from Germany. In conclusion, the continuous-time latent-state approach of MMMPPs offers a natural framework to analyze the evolution of patients’ disease activity underlying claims data by jointly modeling observations and their informative time points.

**ACKNOWLEDGEMENTS**

We thank Roland Langrock for stimulating discussions and helpful comments as well as Rebecca Louise Hilder for valuable feedback on a first draft of the manuscript. We are also grateful to two anonymous reviewers for their insightful and very useful feedback that helped us improve this article. This research was funded by the German Research Foundation (DFG) as part of the SFB TRR 212 (NC\textsuperscript{3}) – Projektnummer 316099922. Open Access funding enabled and organized by Projekt DEAL.

**CONFLICT OF INTEREST**

The authors declare no potential conflict of interests.

**DATA AVAILABILITY STATEMENT**

The data for the case study cannot be shared due to privacy, but artificially simulated data based on the case study results and structured exactly as the real dataset are available for illustration in Appendix S3.

**ORCID**

Sina Mews https://orcid.org/0000-0003-1138-3185
Svenja Elkenkamp https://orcid.org/0000-0001-8704-1208

**REFERENCES**

1. Dutta EK, Kumar S, Venkatachalam S, Downey LE, Albert S. An analysis of government-sponsored health insurance enrolment and claims data from Meghalaya: insights into the provision of health care in north East India. *PLoS One*. 2022;17(6):e0268858.
2. Nerius M, Fink A, Doblhammer G. Parkinson’s disease in Germany: prevalence and incidence based on health claims data. *Acta Neurol Scand*. 2017;136(5):386-392.
3. Min X, Yu B, Wang F. Predictive modeling of the hospital readmission risk from patients’ claims data using machine learning: a case study on COPD. *Sci Rep*. 2019;9:2362.
4. Nielsen SS, Warden MN, Camacho-Soto A, Willis AW, Wright BA, Racette BA. A predictive model to identify Parkinson disease from administrative claims data. *Neurology*. 2017;89(14):1448-1456.
5. Christensen T, Frandsen A, Glazier S, Humpherys J, Kartchner D. Machine learning methods for disease prediction with claims data. In: 2018 IEEE International Conference on Healthcare Informatics (ICHI); 2018; New York, NY:467-4674.
6. Hossain ME, Khan A, Moni MA, Uddin S. Use of electronic health data for disease prediction: a comprehensive literature review. *IEEE/ACM Trans Comput Biol Bioinform*. 2019;18(2):745-758.
7. Ploner T, Heß S, Grum M, Drewe-Boss P, Walker J. Using gradient boosting with stability selection on health insurance claims data to identify disease trajectories in chronic obstructive pulmonary disease. *Stat Methods Med Res*. 2020;29(12):3684-3694.
8. Bureau A, Shiboski S, Hughes JP. Applications of continuous time hidden Markov models to the study of miscategorised disease outcomes. *Stat Med*. 2003;22(3):441-462.
9. Jackson CH, Sharples LD, Thompson SG, Duffy SW, Couto E. Multistate Markov models for disease progression with classification error. *Statistician*. 2003;52(2):193-209.
10. Altman RM, Petkau AJ. Application of hidden Markov models to multiple sclerosis lesion count data. *Stat Med*. 2005;24(15):2335-2344.
11. Amoros R, King R, Toyoda H, Kumada T, Johnson PJ, Bird TG. A continuous-time hidden Markov model for cancer surveillance using serum biomarkers with application to hepatocellular carcinoma. *Metron*. 2019;77:67-86.
12. Scott SL, Smyth P. The Markov modulated Poisson process and Markov Poisson cascade with applications to web traffic modeling. In: Bernardo JM, Bayarri MJ, Berger JO, et al., eds. *Bayesian Statistics*. Oxford, UK: Oxford University Press; 2003:671-680.
13. Langrock R, Borchers DL, Skaug HJ. Markov-modulated nonhomogeneous Poisson processes for modeling detections in surveys of marine mammal abundance. *J Am Stat Assoc*. 2013;108(503):840-851.
14. Lu S. Markov modulated Poisson process associated with state-dependent marks and its applications to the deep earthquakes. *Ann Inst Stat Math*. 2012;64(1):87-106.
15. Choquet R. Markov-modulated Poisson processes as a new framework for analysing capture-recapture data. *Methods Ecol Evol*. 2018;9(4):929-935.
16. Pan J, Rao V, Agarwal P, Gelfand A. Markov-modulated marked Poisson processes for check-in data. In: Balcan MF, Weinberger KQ, eds. Proceedings of The 33rd International Conference on Machine Learning. New York, NY; 2016:2244-2253.

17. Hatt T, Feuerriegel S. Early detection of user exits from clickstream data: a Markov modulated marked point process model. In: Huang Y, King I, Liu TY, van Steen M, eds. WWW’20: Proceedings of The Web Conference 2020. Taipei, Taiwan; 2020:1671-1681.

18. Lange JM, Hubbard RA, Inoue LY, Minin VN. A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. Biometrics. 2015;71(1):90-101.

19. Alaa AM, Hu S, van der Schaar M. Learning from clinical judgments: semi-Markov-modulated marked Hawkes processes for risk prognosis. In: Precup D, Teh YW, eds. Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia; 2017:60-69.

20. Gasparini A, Abrams KR, Barrett JK, et al. Mixed-effects models for health care longitudinal data with an informative visiting process: a Monte Carlo simulation study. Stat Neerl. 2020;74(1):5-23.

21. Su L, Cheng Y, Pereira DJ, Powell JJ. Modelling disease progression with multi-level electronic health records data and informative observation times: an application to treating iron deficiency anaemia in primary care of the UK. arXiv preprint arXiv:2107.13956; 2021.

22. Zucchini W, MacDonald IL, Langrock R. Hidden Markov Models for Time Series: an Introduction Using R. Boca Raton, FL: Chapman & Hall/CRC; 2016.

23. Moler C, Van Loan C. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev. 2003;45(1):3-49.

24. Fouchet D, Santin-Janin H, Sauvage F, Yoccoz NG, Pontier D. An R package for analysing survival using continuous-time open capture-recapture models. Methods Ecol Evol. 2016;7(5):518-528.

25. Mews S, Langrock R, King R, Quick N. Multi-state capture-recapture models for irregularly sampled data. Ann Appl Stat. 2022;16(2):982-998.

26. Luo Y, Stephens DA, Verma A, Buckner AJ, Su L, Cheng Y, Pereira DJ, Powell JJ. Modelling disease progression with multi-level electronic health records data and informative observation times: an application to treating iron deficiency anaemia in primary care of the UK. arXiv preprint arXiv:2107.13956; 2021.

27. Akmatov MK, Stoffen A, Holstiege J, Bätzing J. Chronic obstructive pulmonary disease (COPD) in ambulatory care in Germany—temporal trends and small-area variations. Central Research Institute for Ambulatory Health Care in Germany (Zi). Versorgungsatlas Report 19/06; 2019.

28. Quan H, Li B, Couris CM, et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. Am J Epidemiol. 2011;173(6):676-682.

29. Bannay A, Chaignot C, Blotière PO, et al. The best use of the Charlson comorbidity index with electronic health care database to predict mortality. Med Care. 2016;54(2):188-194.

30. Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inform Theory. 1967;13(2):260-269.

31. Forney GD. The Viterbi algorithm. Proc IEEE. 1973;61(3):268-278.

32. Höhn A, Gampe J, Lindahl-Jacobsen R, Christensen K, Oksuyzan A. Do men avoid seeking medical advice? A register-based analysis of gender-specific changes in primary healthcare use after first hospitalisation at ages 60+ in Denmark. J Epidemiol Community Health. 2020;74(7):573-579.

33. Simon-Tuval T, Scharf SM, Maimon N, Bernhard-Scharf BJ, Reuveni H, Tarasiuk A. Determinants of elevated healthcare utilization in patients with COPD. Respir Res. 2011;12:7.

34. Hutchinson AF, Graco M, Rasekaba TM, Parikh S, Berlowitz DJ, Lim WK. Relationship between health-related quality of life, comorbidities and acute health care utilisation, in adults with chronic conditions. Health Qual Life Outcomes. 2015;13:69.

35. Gruger J, Kay R, Schumacher M. The validity of inferences based on incomplete observations in disease state models. Biometrics. 1991;47(2):595-605.

36. Pullenayegum EM, Lim LS. Longitudinal data subject to irregular observation: a review of methods with a focus on visit processes, assumptions, and study design. Stat Methods Med Res. 2016;25(6):2992-3014.

37. Farzanfar D, Abumuamar A, Kim J, Sirotich E, Wang Y, Pullenayegum E. Longitudinal studies that use data collected as part of usual care risk reporting biased results: a systematic review. BMC Med Res Methodol. 2017;17:133.

38. Chen B, YG Y, Cook RJ. Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process. Stat Med. 2010;29(11):1175-1189.

39. Sweeting M, Farewell V, De Angelis D. Multi-state Markov models for disease progression in the presence of informative examination times: an application to hepatitis C. Stat Med. 2010;29(11):1161-1174.

40. Chen B, Zhou XH. Non-homogeneous Markov process models with informative observations with an application to Alzheimer's disease. Biom J. 2011;53(3):444-463.

41. Chen B, Zhou XH. A correlated random effects model for non-homogeneous Markov processes with nonignorable missingness. J Multivar Anal. 2013;117:1-13.

42. Pohle J, Langrock R, vFMB, Schmidt NM. Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement. J Agric Biol Environ Stat. 2017;22(3):270-293.

43. Cook RJ, Lawless JF. Statistical issues in modeling chronic disease in cohort studies. Stat Biosci. 2014;6(1):127-161.

44. Hui-Min WG, Chang SH, Hsiu-Hsi CT. A Bayesian random-effects Markov model for tumor progression in women with a family history of breast cancer. Biometrics. 2008;64(4):1231-1237.

45. Altman RM. Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. J Am Stat Assoc. 2007;102(477):201-210.
46. Chen PL, Tien HC. Semi-Markov models for multistate data analysis with periodic observations. Commun Stat - Theory Methods. 2004;33(3):475-486.
47. Alaa AM, Van Der Schaar M. A hidden absorbing semi-Markov model for informatively censored temporal data: learning and inference. J Mach Learn Res. 2018;19(1):108-169.
48. Hubbard R, Lange J, Zhang Y, Salim B, Stroud J, Inoue L. Using semi-Markov processes to study timeliness and tests used in the diagnostic evaluation of suspected breast cancer. Stat Med. 2016;35(27):4980-4993.

SUPPORTING INFORMATION
Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Mews S, Surmann B, Hasemann L, Elkenkamp S. Markov-modulated marked Poisson processes for modeling disease dynamics based on medical claims data. Statistics in Medicine. 2023;42(21):3804-3815. doi: 10.1002/sim.9832