MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190358

Modeling Chronic Obstructive Pulmonary Disease Progression Using Continuous-Time Hidden Markov Models

Guido Antonio Powell[a], Aman Verma[a], Yu Luo[a], David Stephens[b], David Buckeridge[a]

[a] Department of Epidemiology and Biostatistics, McGill University, Montreal, Quebec, Canada,
[b] Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada

Abstract

Understanding the progression of chronic diseases, such as chronic obstructive pulmonary disease (COPD), is important to inform early diagnosis, personalized care, and health system management. Data from clinical and administrative systems have the potential to advance this understanding, but traditional methods for modelling disease progression are not well-suited to analyzing data collected at irregular intervals, such as when a patient interacts with a healthcare system. We applied a continuous-time hidden Markov model to irregularly-spaced healthcare utilization events and patient-level characteristics in order to analyze the progression through discrete states of 76,888 patients with COPD. A 4-state model allowed classification of patients into interpretable states of disease progression and generated insights about the role of comorbidities, such as cardiovascular diseases, in accelerating severe trajectories. These results can improve the understanding of the evolution of COPD and point to new hypotheses about chronic disease management and comorbidity.

Keywords:

Latent multi-state models, chronic obstructive pulmonary disease, disease progression.

Introduction

Chronic diseases typically progress slowly over many years. For example, patients with chronic obstructive pulmonary disease (COPD) may progress from mild to very severe disease over the span of more than a decade [1]. Understanding how, and to the extent possible why, patients progress through their disease is crucial for improving disease detection, management and prognosis. For example, inappropriate care at early stages of the disease can accelerate progression to more severe stages.[2,3] At a clinical level, the GOLD criteria define stages of COPD disease (based on lung function assessments) and identify appropriate care by disease stage.[4] However, in the absence of accessible clinical and physiological data for the whole COPD population, health system managers are unaware of the distribution of disease severity and cannot tailor interventions to alter their trajectories.

Fortunately, the data needed to model disease progression exist in the form of the vast amount of individual patient data stored in longitudinal health records, such as healthcare administrative databases. Despite their value, these data are not direct observations of the underlying processes that drives disease progression, but instead they record indirect measures such as healthcare encouters, diagnositc codes and drugs. As a result, inferential methods are needed to model disease progression as a function of these observeables.

Several factors complicate the modeling of disease progression. First, patient data are recorded only when services are delivered, resuling in irregularly-spaced observations that differ in granularity between patients or across one patient's course of illness. Disease trajectories, in terms of the rate of progression and the profile of health service used, will also vary widely across patients, further creating challenges in modeling progression for groups of patients. Moreover, for large segments of patient trajectories, records may be sparse due to infrequent visits, or records may be incomplete (e.g., censored data or care provided outside the health system). Finally, the comorbid conditions experienced by most patients are an important driver of both observed healthcare use and progression in the underlying disease of focus.

Research that has used data from electronical health records to model chronic diseases such as COPD has tended to simplify or collapse the data over time (e.g. 30-day readmission [5]). Other researchers have opted to use regularly spaced data from relatively small and selected clinical cohorts, allowing them to model disease progression as time-to-event [6,7]. In modeling the heterogeneity in COPD progession, researchers have clustered clinical measures cross-sectionally [8] or over-time as latent trajectories from smaller regularly-spaced trial data [9].

Previous disease progression research using mutli-state modeling has typically made use of known disease status or known transition rates [10–13] rather than modeling the latent disease progression as a stochastic process. Recent research by Wang et al. [14] has developed a novel approach to modeling COPD progression using a hidden Markov model (HMM), which assumes a latent Markov process governing an evolution of observed events. These events are defined by latent groupings of COPD-related and comorbidity diagnostic codes (ICD-9) measured in 90-day bins.

To model the heterogeneous, sparse, non-equidistant, and incomplete longitudinal observations in a large cohort of COPD patients we propose using a continuous-time hidden Markov model (CTHMM) under a generalized linear modeling framework with patient-level covariates of comorbidities [15]. Such multi-state models can describe patient status over time as a discrete-time realization of a continuous-time Markov process, while accommodating irregular spacing of observations [16]. We are interested in understanding how a CTHMM can be used to infer latent disease progression from an observed evolution in complex healthcare utilization data.

We use a CTHMM on data from a large cohort of COPD patients to model trajectories of the chronic disease and

examine the role comorbid conditions in the progression. Our objective is to generate hypotheses about the evolution of the illness and how the management of COPD may be improved.

Methods

Data

A cohort of COPD patients was selected from an open, dynamic cohort sampled in 1998 as a 25% random sample drawn from people registered with the provincial health insurance agency (*Régie de l'assurance maladie du Québec*) in the census metropolitan area of Montreal[17]. At the start of every following year, 25% of births and new residents arrived to the area in the past year were sampled to maintain a representative cohort. Patients were followed until they died or moved out of the region. The administrative records included outpatient diagnoses and procedures submitted through billing claims, and procedures and diagnoses from hospital records.

Using established case-definitions based on diagnostic codes [18,19], we enrolled 76,888 COPD patients with an incident event (ICD-9 491x, 492x, 496x; ICD-10 J41-J44) occurring after a minimum of two years at risk and followed patients until December 2014. Comorbid diseases were determined at baseline (upon entry in the larger dynamic cohort) using established case-definitions for 9 diseases: acute myocardial infarction (AMI), asthma, cancer, congestive heart failure (CHF), diabetes, hypertension, ischemic heart disease (IHD), mental illness, and stroke). Patient age and sex were also retained for model covariates.

When a patient has an ecounter with the health system, the data recorded are indirect measures of their underlying disease state. For this study, we used data captured during an outpatient visit with a general practitioner (GP), an outpatient visit with a specialist (SPEC) (coded as respirologist or internist), visits to an emergency department (ED), and hospitalizations (HOSP). Health services used prior to the incident COPD event were not considered.

Model description

To model disease progression within the COPD cohort, we used a continuous-time hidden Markov model (CTHMM). Our model assumes the observed events are generated from hidden states that we interpret as belonging to states of disease severity, by computing the probability of different healthcare utilization events with known levels of severity.[20] The observable data in this CTHMM were healthcare utilization events, which were classifed into four mutually exclusive types (GP, SPEC, ED, HOSP). If two events occured on the same day, then we used the most severe event, where HOSP was the most severe, follwed by ED, SPEC, and GP.

Given an observed event, the probability of each of the four observables was determined by a multinomial model, using GP as the reference event. For each state, there was a set of parameters β that specified the multinomial model of the four healthcare utilization outcomes. Each state's multinomial model was specified as a set of three logistic regression models of the probability of specialist, ED, or hospital visit versus the probability of a GP visit. Each logistic regression model was an intercept-only model, meaning that the probability of each healthcare utilization outcome was the same for any patient in the same state. Fixing the utilization probability within states facilitates the interpretion (or labeling) of different states, strictly based on differences in the severity of patient need across states.

Because the observed values (healthcare utilization events) were irregularly-spaced in time, we modeled the transitions between hidden states in continuous time. Also, as patients may have extended periods without any healthcare visits, we allowed for transitions between states even if there were no observed healthcare utilization events. The transition rate between states was specified by the transition rate matrix Q, whose parameters were fit as part of the model, where each row and column corresponded to a hidden state. If X_t is the state membership at time t, then $q_{ij} = \lim_{\Delta t \to 0} \frac{P(X_{t+\Delta t}=j|X_t=i)}{\Delta t}$ where $i \neq j$. The diagonal entries of Q are computed so that the rows of Q sum to zero. Q was calculated separately for each patient using the vector of their baseline covariates $\mathbf{W}$, such that $\log(q_{ij}) = \mathbf{W}^{\mathrm{T}}\xi_{ij}$ where $i \neq j$.

We fixed the transition matrix to allow transitions only to one adjacent subsequent state (a Markov chain), except in one branching state with two transitions. The parameter set ξ_{ij} for all other state pairs were fixed to zero. By restricting the transitions in this manner we aimed to model a set of alternative disease progression trajectories following an earlier Markov chain progression. We compared the goodness-of-fit of a 4-state model (as an initial state transitioning to a state that branches to two absorbing states) to three models with similar topologies but additional states: 5-states (one state before the branching state), 6-states (one state before each absorbing state), and 7-states (a combination of the 5- and 6-state additions). We compared models using the Bayesian information criteria (BIC), mean posterior probabilities of predicted states, as well as criteria from guidelines on model fitting in HMM concerning the interpretability and usefulness of additional states. [21]

Model fit

Our model required us to fit three sets of parameters. π specified the probability of starting in each state, ξ specified the relationship of the patient's baseline covariates to the transition rate matrix, and β specified the probability of each observable (healthcare utilization) for each state.

To fit our model, we used expectation-maximization (EM) [22], an iterative algorithm in which previous parameter values are used to compute new values. The EM algorihm requires starting values for the model parameters. For π, we specified equal probability of starting in any state, the β parameters were randomly drawn from a normal distribution with a mean of 0 and a standard deviation of 1, and the ξ parameters were drawn from a uniform distribution between 0 and 1, except for the parameter for age, which was scaled by the maximum age in the data. We scaled Δt to be expressed in years.

At each iteration of EM, we used the forward-backward algorithm [23,24], along with the current parameter values for π, β, and ξ, to compute, for each consecutive pair of observations within each patient, the probability of starting and ending in each state pair. From these state pair membership probabilities, we computed the probability of state membership at all observed times for all patients.

We then computed new estimates for each of our parameters. We computed a new value for π using the mean probability of state membership at the first observed time point across all patients, and computed new values for β by fitting a multinomial model for each state, weighting each observation by the probability of state membership at that time. Finally, we computed new values for ξ by using a nested EM procedure [25,26]. We stopped the algorithm when the sum of the

difference in norms between the previous and new parameter values was below 0.05.

Results

Figure 1 (A) describes the β and 5-year, non-diagonal Q coefficients of a 4-state model. A 4-state model was chosen over 5-, 6-, and 7-state Markov chain models despite decreases in BIC in latter models (4-state, 5853716; 5-state, 5767363; 6-state, 5721712; 7-state, 5668360). This decision was based on the 4-state model's higher mean posterior probabilities of predicted states (0.93 vs 5-state, 0.91; 6-state, 0.90; 7-state, 0.87) and its better interpretability of distinct states, allowing for better understanding of predictors of transition in disease progression modeling and for the aim of hypothesis generation. We interpreted state 1 as representing a controlled state with high probability of GP visits (86%). Patients in state 2, interpreted as a state of higher disease severity, had a higher probability of ED visits (51%) and hospitalizations (22%). State 3 was interpreted as a stage of more specialized disease management (40% specialist visits and 41% GP services), compared to state 4 in which patients have a combination of higher GP use (65%) but also higher ED visits (23%).

Patients typically transitioned slowly (16% over 5 years) from state 1 to state 2, while transitions away from state 2 were more accelerated (57% to state 3 and 29% to state 4 over 5 years). The complements of these rates represent the probability of remaining in states 1 (84%) and 2 (14%) over 5 years. Transitions from state 1 to 3 or 4 only occurred through state 2.

Figure 1 (B) shows four illustrative patients trajectories of observed health service use along with predicted states at observation time based on patients' posterior probability. .

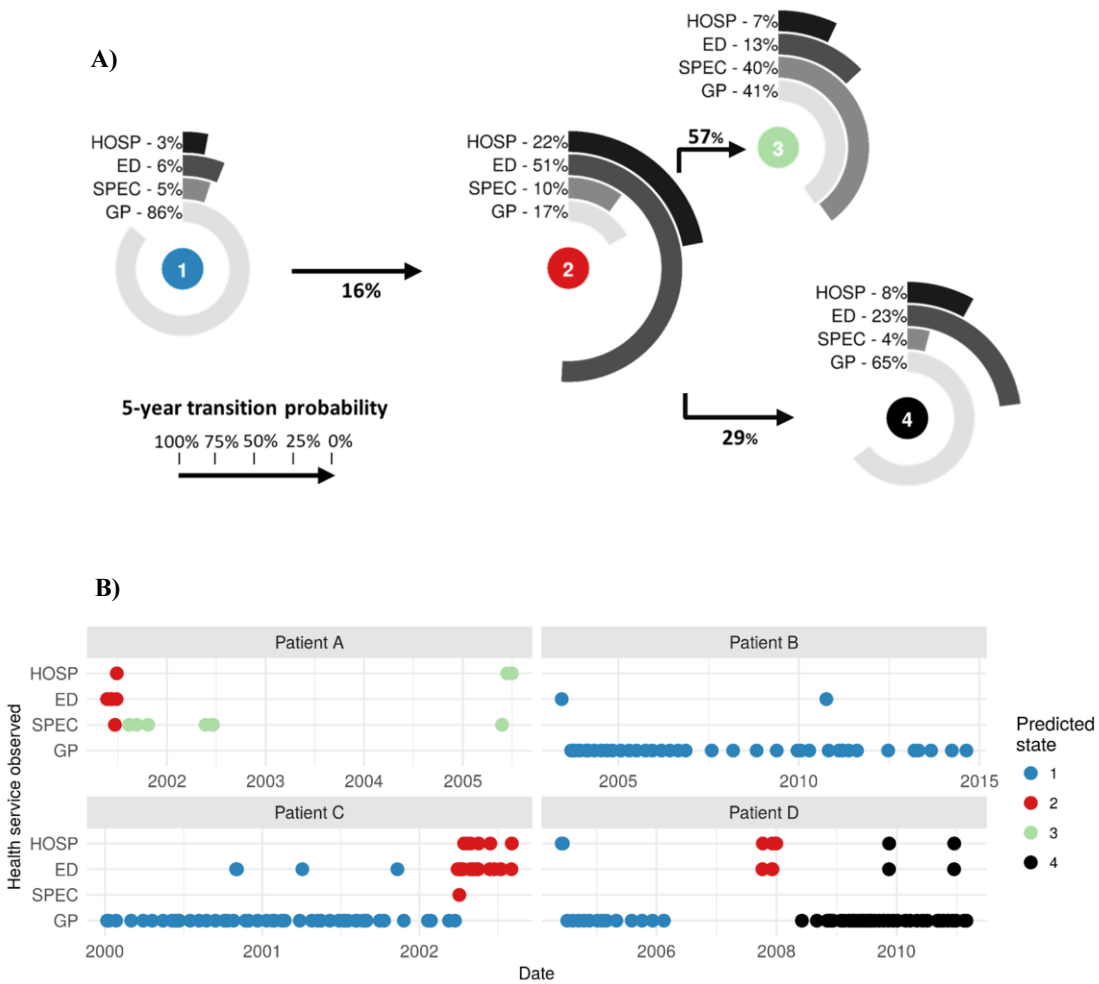

Figure 1 ***A)*** *Diagram of 4-state model of COPD progression based on healthcare utilization. Radial bars represent the probability of each health service (HOSP = hospitalization, ED = emergency department visit, SPEC = specialist visit, GP = general practicioner visit) within a state (labeled 1 to 4), while distance (horizontally) between each state represents the transition probability over 5 years (more distant states represent lower transition probability).* ***B)*** *Observed health service use trajectories of four illustrative patients. Health service types are plotted on separate lines, while patients' predicted state (from1 to 4) at observation time is represented by the dot color*

Table 1 describes the marginal associations between patients' baseline covariates and 5-year transition rates. Noteworthy associations include hypertension, IHD, and stroke increasing the probability of transitioning to and remaining in state 2 (RR ≥ 1.3 and ≥ 1.5, respectively), while cancer is the only covariate associated with higher transition probabilities from state 2 to 3 (1.3). No covariates have large effects on transition probabilities from state 2 to state 4.

Table 1– Risk ratio of transition between states after 5 years by patient characteristics.

Start state	1		2		
End state	1	2	2	3	4
Variable					
Age (10 years)	0.8	1.3	1.0	0.8	1.1
Sex (Male)	0.9	1.2	1.2	0.9	1.0
AMI	1.0	1.0	1.0	1.0	1.0
Asthma	0.9	1.2	1.3	1.1	0.9
Cancer	1.0	1.0	0.9	1.3	0.9
CHF	1.0	1.1	1.2	0.9	1.0
Diabetes	0.9	1.1	0.9	1.0	1.0
Hypertension	0.8	1.8	2.1	0.7	0.9
IHD	0.9	1.4	1.5	0.9	0.9
Mental illness	0.9	1.1	1.1	1.0	1.0
Stroke	0.9	1.3	1.6	0.6	1.1

Discussion

We applied a continuous-time hidden Markov model to healthcare utilization data for a cohort of COPD patients to model their unobserved, latent, disease progression. The model allowed us to learn trajectories through underlying disease states from heterogeneous, sparse and incomplete observations. We found that the model provided interpretable results, allowing us to generate hypotheses about how healthcare utilization evolves at different stages of disease progression.

The discrete states suggest slow progression from a controlled state of the disease (state 1) to one that requires acute care (state 2) and is typically more transient. Patients will mostly transition rapidly from that acute state to one where their disease is managed with primary and specialized care (state 3). However our model specification allowed us to identify an alternative trajectory from state 2 to state 4 where patients are potentially managed inadequately (lower specialist care and higher ED visits than state 3). This pattern suggests possible disparities in patient management (state 3 vs state 4) following hospitalizations and emergency care (state 2) as a window for targeting further investigations into COPD trajectories.

Including patient covariates in the model allowed us to identify predictors of less favourable disease progression patterns. Having a history of different cardiovascular diseases (CHF, hypertension, IHD, and stroke) or of asthma; being older, and being male predicted more severe trajectories, either by precipitating transition towards a state requiring more hospitalizations and ED visits (state 2) or by increasing the probability of remaining in that state. A higher probability of remaining in state 2 typically occurred at the expense of transitions to more specialized care in state 3. We interpret this finding as suggesting that comorbid conditions may increase the risk of repeated acute care in COPD patients and reduce the rate of progression to a state with specialized management. Patients with a history of cancer, however, were more likely to transition to state 3, possibly due to their greater need for specialized treatment. Further analysis can validate these covariates as being modifiers of disease progression by investigating to what extent they predict similar healthcare use patterns outside of a COPD cohort.

These results demonstrate that with little specification beyond model topology (allowed transitions) and number of states, electronic health data can provide new insights and hypotheses on disease progression and management. At a clinical level, such models can improve prognoses by matching patients to typical trajectories of a sub-population. At a health systems level, decision makers can identify health service profiles (e.g. access to specialist care) that are predictive of slower disease progression.

An important aspect of this type of analysis is to weigh multiple criteria in model selection, such as interpretability and relative contribution of additional states [21], especially if the predictors of state transitions are also meaningful. Future research can explore the effect of specifying topologies allowing for transitions back to earlier states, more than one branching state, or higher order Markov chains (states depending on *n* previous states), as well as additional transition covariates such as prescribed drugs and treatments.

Certain limitations of our approach must be considered. Despite our longitudinal focus, covariates of transition probabilities could only be analyzed in terms of their respective baseline measures, a limitation of CTHMMs. Time-varying measures of comorbidities and other covariates would better reflect the evolving impact of patient characteristics on disease progression. In addition, classification of both comorbid status and inclusion in the COPD cohort was based on deterministic case definitions that ignore potential misclassification. An improved approach could learn a probablistic classification of both comorbidities and the modeled disease in a manner similar to how the model learned disease states.

Conclusions

By modeling complex longitudinal observations generated by latent disease states, we have created opportunities for future research in understanding the progression of COPD with potential applications to other chronic diseases.

Acknowledgements

This research was supported by an operating grant from the Canadian Institutes of Health Research.

References

[1] GOLD Executive Committee, Pocket guide to COPD diagnosis, management and prevention: A guide for health care professionals, (2015). http://www.goldcopd.org/uploads/users/files/GOLD_Pocket_2015_Feb18.pdf.

[2] J.J. Soler-Cataluña, M.A. Martínez-García, P. Román Sánchez, E. Salcedo, M. Navarro, and R. Ochando, Severe acute exacerbations and mortality in patients with chronic obstructive pulmonary disease, *Thorax*. **60** (2005) 925–931. doi:10.1136/thx.2005.040527.

[3] S. Spencer, P.M.A. Calverley, P.S. Burge, and P.W. Jones, Impact of preventing exacerbations on deterioration of health status in COPD, *Eur. Respir. J.* **23** (2004) 698–702.
[4] C.F. Vogelmeier, G.J. Criner, F.J. Martinez, A. Anzueto, P.J. Barnes, J. Bourbeau, B.R. Celli, R. Chen, M. Decramer, L.M. Fabbri, P. Frith, D.M.G. Halpin, M.V. López Varela, M. Nishimura, N. Roche, R. Rodriguez-Roisin, D.D. Sin, D. Singh, R. Stockley, J. Vestbo, J.A. Wedzicha, and A. Agustí, Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary, *Am. J. Respir. Crit. Care Med.* **195** (2017) 557–582. doi:10.1164/rccm.201701-0218PP.
[5] R. Sharif, T.M. Parekh, K.S. Pierson, Y.-F. Kuo, and G. Sharma, Predictors of Early Readmission among Patients 40 to 64 Years of Age Hospitalized for Chronic Obstructive Pulmonary Disease, *Annals ATS*. **11** (2014) 685–694. doi:10.1513/AnnalsATS.201310-358OC.
[6] H. Müllerova, D.J. Maselli, N. Locantore, J. Vestbo, J.R. Hurst, J.A. Wedzicha, P. Bakke, A. Agusti, and A. Anzueto, Hospitalized Exacerbations of COPD: Risk Factors and Outcomes in the ECLIPSE Cohort, *Chest*. **147** (2015) 999–1007. doi:10.1378/chest.14-0655.
[7] J. Vestbo, W. Anderson, H.O. Coxson, C. Crim, F. Dawber, L. Edwards, G. Hagan, K. Knobil, D.A. Lomas, W. MacNee, E.K. Silverman, R. Tal-Singer, and on behalf of the E. Investigators, Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE), *European Respiratory Journal*. **31** (2008) 869–873. doi:10.1183/09031936.00111707.
[8] M. Paoletti, G. Camiciottoli, E. Meoni, F. Bigazzi, L. Cestelli, M. Pistolesi, and C. Marchesi, Explorative data analysis techniques and unsupervised clustering methods to support clinical assessment of Chronic Obstructive Pulmonary Disease (COPD) phenotypes, *J Biomed Inform*. **42** (2009) 1013–1021. doi:10.1016/j.jbi.2009.05.008.
[9] D.E. Stull, I. Wiklund, R. Gale, G. Capkun-Niggli, K. Houghton, and P. Jones, Application of latent growth and growth mixture modeling to identify and characterize differential responders to treatment for COPD, *Contemporary Clinical Trials*. **32** (2011) 818–828. doi:10.1016/j.cct.2011.06.004.
[10] P.K. Andersen, and N. Keiding, Multi-state models for event history analysis, *Stat Methods Med Res*. **11** (2002) 91–115.
[11] L. Meira-Machado, J. de Uña-Alvarez, C. Cadarso-Suárez, and P.K. Andersen, Multi-state models for the analysis of time-to-event data, *Stat Methods Med Res*. **18** (2009) 195–222. doi:10.1177/0962280208092301.
[12] J.M. Lange, R.A. Hubbard, L.Y.T. Inoue, and V.N. Minin, A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data: A Joint Model for Multistate Disease Processes and Random Informative Observation Times, *Biometrics*. **71** (2015) 90–101. doi:10.1111/biom.12252.
[13] N. Bartolomeo, P. Trerotoli, and G. Serio, A multistate model to evaluate COPD progression integrating drugs consumption data and hospital databases, *Epidemiology, Biostatistics and Public Health*. **12** (2015). http://ebph.it/article/view/11145 (accessed November 19, 2016).
[14] X. Wang, D. Sontag, and F. Wang, Unsupervised Learning of Disease Progression Models, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2014: pp. 85–94. doi:10.1145/2623330.2623754.
[15] Y. Luo, D.A. Stephens, A. Verma, and D.L. Buckeridge, Latent Multi-State Models For Non-Equidistant Longitudinal Electronic Health Records, (n.d.).
[16] J.D. Kalbfleisch, and J.F. Lawless, The Analysis of Panel Data Under a Markov Assumption, *Journal of the American Statistical Association*. **80** (1985) 863–871. doi:10.2307/2288545.
[17] A. Shaban-Nejad, M. Lavigne, A. Okhmatovskaia, and D.L. Buckeridge, PopHR: a knowledge-based platform to support integration, analysis, and visualization of population health data, *Ann. N.Y. Acad. Sci.* (2016). doi:10.1111/nyas.13271.
[18] C. Blais, S. Jean, C. Sirois, L. Rochette, C. Plante, I. Larocque, M. Doucet, G. Ruel, M. Simard, P. Gamache, D. Hamel, D. St-Laurent, and V. Emond, Quebec Integrated Chronic Disease Surveillance System (QICDSS), an innovative approach, *Chronic Dis Inj Can*. **34** (2014) 226–235.
[19] L. Lix, J. Ayles, S. Bartholomew, C. Cooke, J. Ellison, V. Emond, N. Hamm, H. Hannah, S. Jean, S. LeBlanc, J.M. Paterson, C. Pelletier, K. Phillips, R. Puchtinger, K. Reimer, C. Robitaille, M. Smith, L. Svenson, K. Tu, L. VanTil, S. Waits, and L. Pelletier, The Canadian Chronic Disease Surveillance System: A model for collaborative surveillance, *1*. **3** (2018). doi:10.23889/ijpds.v3i3.433.
[20] A.S. Gershon, G.C. Mecredy, S.D. Aaron, P.G. Camp, K. Tu, P. Hernandez, and T. To, Development of quality indicators for chronic obstructive pulmonary disease (COPD): A modified RAND appropriateness method, *Canadian Journal of Respiratory, Critical Care, and Sleep Medicine*. **3** (2019) 30–38. doi:10.1080/24745332.2018.1476030.
[21] J. Pohle, R. Langrock, F.M. van Beest, and N.M. Schmidt, Selecting the Number of States in Hidden Markov Models: Pragmatic Solutions Illustrated Using Animal Movement, *JABES*. **22** (2017) 270–293. doi:10.1007/s13253-017-0283-8.
[22] A.P. Dempster, N.M. Laird, and D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*. **39** (1977) 1–38.
[23] L.E. Baum, and J.A. Eagon, An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology, *Bull. Amer. Math. Soc.* **73** (1967) 360–363.
[24] L.E. Baum, and G.R. Sell, Growth transformations for functions on manifolds., *Pacific J. Math.* **27** (1968) 211–227.
[25] M. Bladt, and M. Sørensen, Statistical Inference for Discretely Observed Markov Jump Processes, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. **67** (2005) 395–410.
[26] P. Metzner, I. Horenko, and C. Schütte, Generator estimation of Markov jump processes based on incomplete observations nonequidistant in time, *Phys Rev E Stat Nonlin Soft Matter Phys*. **76** (2007) 066702. doi:10.1103/PhysRevE.76.066702.

Address for correspondence

Guido Antonio Powell, email: guido.powell@mail.mcgill.ca