مجله تخصصی اپیدمیولوژی ایران؛ تابستان ۱۳۹۷، دوره ۱٤، شماره ۲: صفحات: ۱۳۵-۱۲٦.

مقاله پژوهشی

استفاده از مدل مارکوف پنهان در پیش‌بینی موارد جدید سل در استان همدان بر اساس اطلاعات موارد ثبت شده طی سال‌های ۹۴-۱۳۸۴

ملیحه صفری[۱]، مجید صادقی‌فر[۲]، قدرت‌اله روشنایی[۳]، علی ظهیری[٤]

[۱] دانشجوی دکترای آمار زیستی، گروه آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی همدان، همدان، ایران

[۲] استادیار آمار، گروه آمار دانشکده علوم، دانشگاه بوعلی سینا همدان، همدان، ایران

[۳] دانشیار آمار زیستی، مرکز تحقیقات مدل‌سازی بیماری‌های غیر واگیر، گروه آمار زیستی و اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی همدان، همدان، ایران

[٤] کارشناس بیماری‌های معاونت بهداشتی، دانشگاه علوم پزشکی همدان، همدان، ایران

نویسنده رابط: قدرت‌اله روشنایی، نشانی: همدان، خیابان شهید فهمیده روبروی پارک مردم، دانشگاه علوم پزشکی همدان، دانشکده بهداشت، گروه آمارزیستی و اپیدمیولوژی، تلفن:۰۸۱-۳۸۳۸۰۰۹۰، پست الکترونیک: gh.roshanaei@umsha.ac.ir

تاریخ دریافت: ۹۶/۰۸/۲۵؛ پذیرش: ۹۷/۰۲/۰۱

مقدمه و اهداف: سل یک بیماری باکتریایی مزمن و به‌عنوان یک عامل مهم ابتلا و مرگ‌ومیر مطرح بوده و در اثر مجموعه‌ای از مایکوباکتریوم‌های سلی ایجاد می‌شود. آگاهی از بروز و تعداد موارد جدید این بیماری اطلاعات ارزشمندی را برای بازنگری برنامه‌ها و شاخص‌های توسعه فراهم می‌کند. مدل‌های سری زمانی و رگرسیون از مدل‌های متداول برای پیش‌بینی بوده، اما مستلزم پیش‌فرض‌هایی هستند. هدف این مطالعه پیش‌بینی موارد جدید بیماری با استفاده از مدل مارکوف پنهان است.

روش کار: داده‌های این مطالعه تعداد موارد جدید سل در استان همدان به‌صورت ماهانه طی سال‌های ۹۴-۱۳۸۴ که توسط مرکز بهداشت استان همدان شناسایی شد؛ بود. در این مطالعه پیش‌بینی موارد جدید سل برای ۲۴ ماه آینده با استفاده از مدل مارکوف پنهان و با نرم‌افزار R بسته مارکوف پنهان انجام شد.

یافته‌ها: بر اساس معیار برازش مدل، یک مدل مارکوف با دو حالت بهترین برازش را به داده‌ها داشت یعنی داده‌های این مطالعه آمیخته‌ای از دو توزیع پواسن با پارامتر میانگین تعداد رخداد ۵/۹۶ و ۱۰/۲ هستند. هم‌چنین یافته‌های پیش‌بینی بر اساس مدل مارکوف پنهان، تعداد موارد جدید سل طی ۲۴ ماه آینده را بین ۹-۸ مورد جدید پیش‌بینی کرد.

نتیجه‌گیری: مدل‌های مارکوف پنهان مناسب‌ترین مدل پیش‌بینی با استفاده از زنجیر مارکوف است که علاوه بر شناسایی مدل مناسب، قادر است ماتریس احتمال انتقال بین حالات مختلف بیماری را تعیین کند تا این احتمالات به پزشکان در پیش‌بینی مراحل آتی بیماری و انجام اقدامات پیش‌گیرانه پیش از ورود به مراحل پیشرفته یاری نماید.

واژگان کلیدی: سل، مدل پنهان مارکوف، پیش‌بینی، همدان

مقدمه

سل یا توبرکلوز یک بیماری عفونی است که بیش‌تر توسط مایکوباکتریوم توبرکلوزیس (TB) ایجاد می‌شود. سل در سال ۱۹۹۰ میلادی در رده هفتم بوده و پیش‌بینی شده که در سال ۲۰۲۰ میلادی هم‌چنان در این رده باقی بماند (۱). با توجه به اولویت‌های کنترلی سل، سازمان جهانی بهداشت اهدافی را برای کنترل سل مشخص کرده است که مهم‌ترین آن عبارت است از این که شیوع سل تا سال ۲۰۱۵ میلادی در جهان به ۵۰ درصد میزان آن در سال ۱۹۹۰ میلادی برسد و تا سال ۲۰۵۰ میلادی میزان مرگ ناشی از سل فعال به یک نفر در یک میلیون نفر کاهش یابد. این در حالی است که این سازمان در سال ۲۰۰۹ میلادی گزارش کرد که ۹/۴ میلیون نفر در دنیا به سل آلوده هستند که نسبت به سال ۲۰۰۷ میلادی، ۱۳۰۰۰۰ نفر افزایش داشته است (۲). برآورد شده است که ۹ میلیون مورد جدید مبتلا به سل و دو میلیون مرگ به دلیل این بیماری در هر سال رخ می‌دهد. به‌علاوه ۲/۵ درصد بار بیماری در جهان به دلیل سل است. در ایران اگرچه اجرای برنامه ملی سل (NTP)[۱] باعث کنترل سل در سال‌های اخیر شده است و تعداد موارد مرگ‌ومیر در دو دهه اخیر کاهش یافته، اما این بیماری تهدید جدی برای سلامت عمومی است (۵-۳). مطالعه‌ها نشان داده‌اند که تعداد موارد این

[۱] National Tuberculosis Program; NTP

بیماری طبق مدل‌های سری‌های زمانی دارای الگوی متفاوت در کشورهای مختلف است به‌ویژه در برخی از کشورها بیش‌ترین فراوانی موارد جدید بیماری در انتهای فصل زمستان و شروع فصل بهار گزارش شده است (۶،۷). تشخیص زودرس طغیان بیماری یکی از اهداف اصلی سامانه‌های نظارتی و پایش است و به همین دلیل است که در دو دهه گذشته روش‌های پایش سنتی به‌وسیله سامانه‌های پایشی زیستی[1] که هدف آن‌ها کاهش زمان تا و گزارش طغیان بیماری است، جایگزین شده است (۸). سامانه جدید پایش، سامانه هشدار سریع اپیدمی را از طریق نظارت براساس داده‌های سری زمانی تعداد موارد بروز بیماری که به صورت ماهیانه یا هفتگی جمع‌آوری شده ارایه می‌دهد (۹). روش‌های آماری که برای تشخیص زودرس و به‌موقع طغیان به‌کار می‌روند، براساس داده‌های پایش مبتنی است. این روش‌ها شامل روش‌های رگرسیونی، سری‌های زمانی، کنترل فرایند آماری و روش‌های بیزی که برای نظارت بر پایش اپیدمیولوژیک بیماری‌های عفونی به‌کار می‌روند (۱۰). درحالی که بیماری‌های عفونی و مسری درون یکی از دو مرحله‌ی اپیدمی و غیراپیدمی قرار می‌گیرند (۱۱). در سال‌های اخیر مدل‌های ریاضی برای پیش‌بینی رخدادهایی به‌کار می‌رود. مدل‌های مارکوف، انتقال‌های وابسته به زمان بین حالت‌های یک سامانه هم‌چون ابتلا به عفونت مایکوباکتریوم توبرکلوزیس و پیشرفت تا مرحله سل فعال را مدل‌بندی می‌کنند. مدل مارکوف پنهان (HMM)[2] نیز می‌تواند انتقال از حالتی به حالت دیگر که در بیماری‌های عفونی به صورت مستقیم قابل مشاهده نیست را آشکار نماید. هم‌چنین این مدل‌ها می‌توانند برای پیش‌بینی رخدادهای بالینی که تاکنون رخ نداده‌اند، استفاده شوند (۱۲).

واتکینز و همکاران داده‌های سری زمانی مربوط به نرخ بیماری شبه آنفلوانزا و تعداد موارد فلج اطفال را مدل‌بندی نموده و نشان دادند که HMM در مدل‌بندی داده‌هایی که به منظور پایش روزمره بیماری‌ها به کار می‌روند، بسیار توانمند است. با این وجود این مدل‌ها به ندرت در سامانه‌های سلامت عمومی به‌کار رفته‌اند (۱۳). هدف اصلی سری زمانی پیش‌بینی برای آینده است، در صورتی که داده‌ها مربوط به به یک سری زمانی از شمارش‌ها باشد نمی‌توان از مدل‌های استاندارد سری زمانی هم‌چون مدل‌های ARIMA استفاده کرد، زیرا این مدل‌ها مبتنی بر فرض نرمال بودن توزیع داده‌ها است که این فرض برای داده‌های شمارشی معمولاً برقرار نیست. یکی از مدل‌های معمول برای داده‌های شمارشی توزیع پواسن است و در توزیع پواسن واریانس با میانگین برابر است، اما بیش‌تر داده‌های شمارشی این فرض برقرار نیست. بنابراین استفاده از مدل پواسن برای مدل‌سازی این نوع از داده‌های شمارشی نامناسب است. بیش‌تر داده‌های شمارشی ممکن است در برخی دوره‌های زمانی دارای نرخ کم رخداد و در برخی دوره‌های زمانی دارای نرخ وقوع نسبتاً بالایی باشند. بنابراین نیاز به مدلی است که توزیع احتمالی هر مشاهده وابسته به حالت‌های پنهان یا غیرقابل مشاهده از زنجیر مارکوف باشد تا هر دو شرایط بیش‌پراکنشی و وابستگی سریالی داده‌ها را به حساب آورد. به مدل‌هایی که این دو شرط داده‌های شمارشی را درنظر می‌گیرند مدل مارکوف پنهان گفته می‌شود (۱۶-۱۴). به‌نظر می‌رسد استفاده از مدل‌های آمیخته در این حالت بر مدل‌هایی که از یک توزیع واحد برای تمام مشاهده‌ها استفاده می‌کنند، برتری دارد. ایده‌ی اصلی استفاده از مدل مارکوف پنهان در پایش اپیدمیولوژی بیماری‌های عفونی و مسری در سال ۱۹۹۹ میلادی ارایه شد (۱۷).

مینلی و همکاران (۲۰۱۳) از HMM برای پیش‌بینی روند پیشرفت سرطان ریه استفاده کردند (۱۸). ویمالا و همکاران (۲۰۱۴) HMM را برای تشخیص و طبقه‌بندی تعداد سیگنال‌های نوار قلب به‌کار بردند (۱۹). رافعی و همکاران (۲۰۱۵) در پیش بینی بروز سل ریوی اسمیر مثبت از این مدل استفاده کردند (۲۰). مددی‌زاده و همکاران (۲۰۱۶) با استفاده از مدل HMM به پیش‌بینی و تشخیص حالات مختلف بیماری کبدی پرداختند (۲۱).

تاکنون در بسیاری از مطالعه‌ها برای پیش‌بینی بروز سل از مدل‌های سری زمانی و رگرسیونی استفاده شده است و مطالعه‌های کمی از مدل‌های مارکوف استفاده کرده‌اند (۳۰-۲۲) و با توجه به این‌که داده‌های سل شمارشی بوده و دارای توزیع نرمال نیستند و هم‌چنین دارای مشکل بیش‌پراکنش نیز باشند، بنابراین این مطالعه به پیش‌بینی موارد جدید بیماران مبتلا به سل در استان همدان برای ۲۴ ماه آینده با استفاده از HMM می‌پردازد.

روش کار

نوع مطالعه و داده‌ها

این مطالعه یک مطالعه توصیفی- تحلیلی است که در آن تمامی موارد جدید مبتلا به سل طی سال‌های ۹۴-۱۳۸۴ از اطلاعات موجود در برنامه نرم افزاری TB Register مربوط به بیماران مبتلا به سل شناسایی شده استان همدان مورد بررسی قرار گرفت.

[1] Biosurveillance
[2] HiddenMarkov Model; HMM

بنابراین تعداد بیماران مبتلا به سل شناسایی شده جدید و بومی استان همدان به تفکیک ماه استخراج شد. تعداد کل بیماران مبتلا به سل در این مطالعه طی ۱۱ سال برابر ۱۲۲۴ بیمار بود.

معرفی مدل آماری مورد استفاده:

HMM یک ابزار آماری برای برازش یک توزیع آمیخته در دنباله‌ای از داده‌های وابسته است. این مدل در تحلیل سیگنال الکتروکاردیوگرافی، تحلیل فراوانی موارد صرع، تحلیل دنباله DNA استفاده شده است (۳۱). یک HMM شامل فرایند زمان گسسته دومتغیره شبیه $\{S_t, Y_t\}_{t\geq 1}$ است که در آن $\{S_t\}$ یک زنجیر مارکوف غیرقابل مشاهده و $\{Y_t\}$ به شرط $\{S_t\}$ دنباله‌ای از متغیرهای تصادفی مستقل است به‌طوری‌که توزیع شرطی Y_t تنها به S_t وابسته است. دنباله‌های $\{Y_t\}$ و$\{S_t\}$ به‌ترتی‌ب دنباله مشاهده‌ها و دنباله فضای وضعیت (حالت) گویند (۲۳). فرض کنید S_t، t=1,2,...,n، نشان‌دهنده زنجیر مارکوف مرتبه اول با یکی از مقادیر 1,2,...,m با ماتریس احتمال انتقال $\Gamma = (\alpha_{ij})_{m*m}$ و توزیع احتمال اولیه $\pi = (\pi_1, \dots, \pi_m)$ باشد که α_{ij} احتمال انتقال از حالت i به حالت j و π_i احتمال آمدن مشاهدات از حالت‌های مختلف را نشان می‌دهد که در آن

$$\pi_i = P(S_t = j\alpha_{ij} = P(S_t = j|S_{t-1} = i)\ i, 理 = 1, \dots, m; t = 1, \dots, n), i = 1, \dots, n.$$

به علاوه توزیع شرطی Y_t به شرط $S_t = i$از توزیع پارامتری $f_i(y_t; \theta_i)$ پیروی می‌کند و θ_i بردار پارامترهای نامعلوم است که نیاز است در این روش براساس اطلاعات موجود این پارامترها (شامل مقادیر اولیه π، ماتریس احتمال انتقال Γ و پارامترهای توزیع $\boldsymbol{\theta}_i$) برآورد شوند. بعد از برآورد پارامترها حالت‌های پنهانی که داده‌ها را تولید کرده‌اند شناسایی می شوند.

همان‌طور که گفته شد HMM قصد دارد تعداد حالت‌هایی که داده‌ها از آن تولید می‌شوند و احتمال انتقال بین حالت‌های مختلف را کشف نماید. با توجه به این‌که برای وقوع سل تعداد حالت‌ها یا فازهای متفاوتی از شدت عفونت‌زایی بیماری وجود دارد که شامل یک حالت معمول با شدت عفونت‌زایی خفیف و چندین حالت غیرمعمول با عفونت‌زایی شدید است که ممکن است تعداد این حالات غیرمعمول از جامعه‌ای به جامعه دیگر متفاوت باشد که حالت معمول متناظر با بروز بیماری در حد انتظار و حالت غیرمعمول متناظر با بروز بیماری بیش از حد انتظار است. در این مطالعه y_t تعداد موارد گزارش شده ماهیانه مبتلا به سل است که از چندین توزیع آمیخته تولید شده‌اند که قرار است HMM ارایه شده این حالت‌ها را شناسایی نماید و S_t حالت‌های مختلف بیماری را نشان می‌دهد. با توجه به این‌که داده‌های این مطالعه شامل داده‌های شمارشی گسسته است، بنابراین انتظار می‌رود داده‌ها، آمیخته‌ای از چند توزیع پواسن با پارامترهای مختلف باشند. بنابراین HMM از توزیع شرطی

$Y_t = y|\ S_t = i\ i=1,2,...,m$

به‌دست می‌آید که هدف این مطالعه به‌دست آوردن حالت بیماری در هر ماه براساس تعداد موارد مشاهده شده در آن ماه است که این توزیع احتمال به صورت زیر تعریف می‌شود.

$$P(Y_t = y|\ S_t = i) = \frac{e^{-\lambda_{it}}\lambda_{it}^{y}}{y!} \quad i = 1,2, \dots, m$$

که در آن λ_{it} برابر تعداد مورد انتظار بیماران در هر حالت است.

روش برآورد و نرم‌افزار مورد استفاده در این مطالعه

برای تعیین لامبدا (تعداد موارد مورد انتظار) در هر مرحله، چندین مدل با تعداد حالات مختلف برازش شد و با آزمون‌های نیکویی برازش تعداد حالت بهینه انتخاب شد. برازش HMM به داده‌ها مستلزم برآورد پارامترها (احتمالات انتقال، مقادیر اولیه و پارامترهای توزیع) است. از روش برآورد بیشینه درست‌نمایی برای برآورد پارامترها استفاده شد. سپس محتمل‌ترین دنباله از مراحل پنهان که داده‌ها از آن تولید شده‌اند بر اساس معیارهای AIC و BIC شناسایی شد و برای پیش‌بی‌نی تعداد موارد جدید سل برای ۲۴ ماه آینده از نرم‌افزار R بسته مارکوف پنهان[1] استفاده شد.

[1] Hidden Markov Package

یافته‌ها

ابتدا تعداد موارد جدید و میزان بروز سل به تفکیک سال‌های ۹۴-۱۳۸۴ در جدول شماره ۱ ارایه شده است.

سپس به منظور نمایش تغییرات و وضعیت نوسانات موارد جدید بیماری ثبت شده در هر ماه، سری زمانی تعداد ماهیانه موارد جدید مبتلایان در نمودار شماره ۱ نمایش داده شده است.

برای تعیین تعداد حالت‌ها (تعداد توزیع‌هایی که مشاهده‌ها از آن پیروی می‌کنند) مدل‌هایی با تعداد حالات ۴،۳،۲،۱ و ۵ برازش شد. بر اساس معیارهای AIC و BIC، HMM دوحالتی بهترین برازش را به داده‌ها داشت. نتایج برازش مدل نشان داد که داده‌ها در این مطالعه آمیخته‌ای از دو توزیع پواسن با پارامترهای به ترتیب λ۱=۱۰/۲ و λ۲=۵/۹۶ هستند، یعنی بدون استفاده از مدل مارکوف پنهان تنها می‌توان نتیجه گرفت که همه داده‌ها از یک توزیع آمده‌اند، اما همان‌طور که گفته شد می‌توان گفت که موارد جدید مبتلایان به سل در این داده‌ها از دو حالت (فاز) بیماری تشکیل می‌شوند که HMM به کار رفته این موضوع را تأیید کرد.

براساس این مدل، ماتریس احتمال انتقال از هر حالت به حالت دیگر بیماری به‌صورت زیر برآورد شد:

$$\begin{pmatrix} 0/75 & 0/25 \\ 0/07 & 0/93 \end{pmatrix}$$

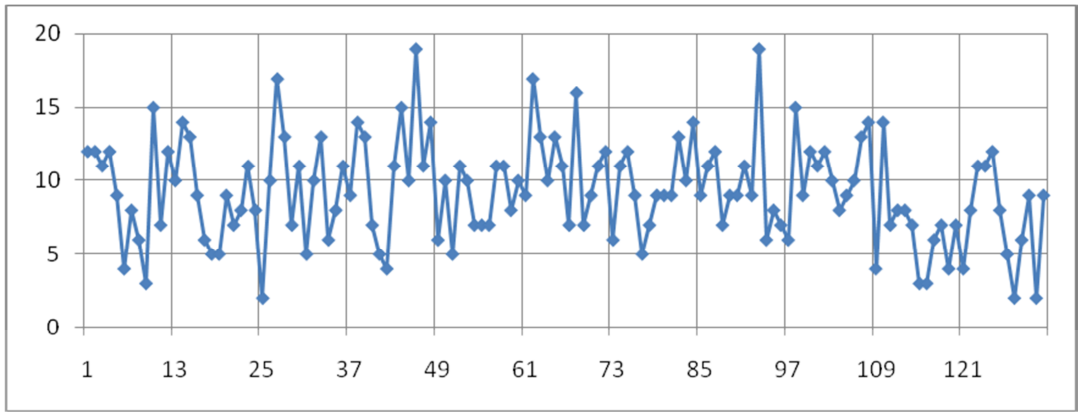
سطر اول ماتریس فوق نشان می‌دهد که یک فرد مبتلا به سل در فاز معمول با احتمال ۷۵ درصد در همان فاز باقی می‌ماند و با احتمال ۲۵ درصد وارد فاز دوم بیماری می‌شوند. سطر دوم هم نشان می‌دهد افراد مبتلایی که در فاز دوم بیماری هستند با احتمال ۷ درصد وارد فاز اول می‌شوند و با احتمال ۹۳ درصد در همان فاز باقی می‌مانند.

همان‌طور که گفته شد یکی از اهداف برازش HMM به داده‌ها، پیش‌بینی تعداد بروز بیماری برای ماه‌های آینده است. جدول شماره ۲ یافته‌های پیش‌بینی HMM دوحالتی برازش یافته به داده‌ها و برخی آمارهای توصیفی را نشان می‌دهد. بر اساس نتایج به‌دست آمده از برازش HMM دو حالتی، تعداد موارد جدید سل طی ۲۴ ماه آینده بین ۹-۸ مورد پیش‌بینی شد.

جدول شماره ۱- فراوانی موارد جدید سل و میزان بروز در هر صدهزار نفر در استان همدان طی سال‌های ۹۴-۱۳۸۴

سال	۱۳۸۴	۱۳۸۵	۱۳۸۶	۱۳۸۷	۱۳۸۸	۱۳۸۹	۱۳۹۰	۱۳۹۱	۱۳۹۲	۱۳۹۳	۱۳۹۴
جمعیت استان*	۱۷۰۰۷۱۹	۱۷۰۳۲۶۷	۱۷۱۷۰۴۱	۱۷۳۶۱۱۳	۱۷۵۵۶۴۵	۱۷۴۷۰۰۰	۱۷۵۸۰۰۰	۱۷۶۸۰۰۰	۱۷۷۷۰۰۰	۱۷۸۶۰۰۰	۱۷۹۵۰۰۰
تعداد موارد جدید سل	۱۲۱	۱۰۷	۱۲۲	۱۳۸	۱۱۲	۱۴۱	۱۲۳	۱۲۳	۱۳۳	۸۲	۸۷
میزان بروز در هر صدهزار نفر	۷/۱	۶/۳	۷/۱	۸	۶/۴	۸/۱	۷	۶/۹۶	۷/۵	۴/۶	۴/۸۵

*تعداد جمعیت در سال‌های ۸۵ و ۹۰ نت‌ی‌جه سرشماری و بق‌ی‌ه سال هابرآورد جمعیت بوده است.



نمودار شماره ۱- سری زمانی فراوانی ماهانه موارد جدید بیماران مبتلا به سل در استان همدان طی سال‌های ۹۴-۱۳۸۴

جدول شماره ۲– احتمالات انتقال به تفکیک ماه و یافته‌های پیش‌بینی تعداد موارد جدید سل با استفاده از مدل مارکوف پنهان دو حالتی برای ۲٤ ماه آینده

ماه		۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲
احتمال انتقال	حالت ۱	۰/۴۶	۰/۳۹	۰/۳۳	۰/۳	۰/۲۷	۰/۲۶	۰/۲۴	۰/۲۳	۰/۲۳	۰/۲۲	۰/۲۲	۰/۲۲
	حالت ۲	۰/۵۴	۰/۶۱	۰/۶۷	۰/۷	۰/۷۳	۰/۷۴	۰/۷۶	۰/۷۷	۰/۷۷	۰/۷۸	۰/۷۸	۰/۷۸
میانگین موارد جدید		۸/۲۱	۸/۵۴	۸/۷۶	۸/۹۲	۹/۰۲	۹/۰۹	۹/۱۴	۹/۱۸	۹/۲	۹/۲۱	۹/۲۲	۹/۲۳
میانه موارد جدید سل		۸	۸	۹	۹	۹	۹	۹	۹	۹	۹	۹	۹
مد موارد جدید سل		۷	۸	۸	۸	۹	۹	۹	۹	۹	۹	۹	۹
فاصله پیش‌بی‌نی	حد پایین	۳	۳	۳	۳	۴	۴	۴	۴	۴	۴	۴	۴
	حد بالا	۱۴	۱۴	۱۴	۱۴	۱۵	۱۵	۱۵	۱۵	۱۵	۱۵	۱۵	۱۵
ماه		**۱۳**	**۱۴**	**۱۵**	**۱۶**	**۱۷**	**۱۸**	**۱۹**	**۲۰**	**۲۱**	**۲۲**	**۲۳**	**۲۴**
احتمال انتقال	حالت ۱	۰/۲۲	۰/۲۲	۰/۲۲	۰/۲۲	۰/۲۲	۰/۲۲	۰/۲۲	۰/۲۲	۰/۲۲	۰/۲۲	۰/۲۲	۰/۲۲
	حالت ۲	۰/۷۸	۰/۷۸	۰/۷۸	۰/۷۸	۰/۷۸	۰/۷۸	۰/۷۸	۰/۷۸	۰/۷۸	۰/۷۸	۰/۷۸	۰/۷۸
میانگین موارد جدید		۹/۲۴	۹/۲۴	۹/۲۴	۹/۲۴	۹/۲۵	۹/۲۵	۹/۲۵	۹/۲۵	۹/۲۵	۹/۲۵	۹/۲۵	۹/۲۵
میانه موارد جدید سل		۹	۹	۹	۹	۹	۹	۹	۹	۹	۹	۹	۹
مد موارد جدید سل		۹	۹	۹	۹	۹	۹	۹	۹	۹	۹	۹	۹
فاصله پیش‌بی‌نی	حد پایین	۴	۴	۴	۴	۴	۴	۴	۴	۴	۴	۴	۴
	حد بالا	۱۵	۱۵	۱۵	۱۵	۱۵	۱۵	۱۵	۱۵	۱۵	۱۵	۱۵	۱۵

بحث

مدل‌های معمول برای پیش‌بینی مستلزم داشتن پیش‌فرض‌هایی است که در صورت نداشتن این پیش‌فرض‌ها یافته‌های حاصل دارای اعتبار لازم نخواهند بود؛ بنابراین نیاز به استفاده از مدل‌هایی است که بتوان با محدودیت‌های کم‌تر یافته‌های لازم را اخذ نمود. HMM یکی از مدل‌های آماری در زمینه تشخیص و طبقه‌بندی است. این مدل به‌خوبی می‌تواند به‌عنوان یک پیشگوی قوی استفاده شود (۱۸،۲۱). در این مطالعه HMM به کار رفته یک مدل مارکوف زمان گسسته بود که با هدف پیش‌بینی موارد جدید سل در استان همدان انجام شد. در صورتی که داده‌های شمارشی دارای توزیع نرمال باشند، می‌توان از سری‌های زمانی برای پیش بینی استفاده کرد، اما با توجه به عدم برقراری فرض نرمال بودن در داده‌های شمارشی، نمی‌توان از مدل‌های استاندارد سری زمانی استفاده کرد. هم‌چنین استفاده از مدل پواسن برای مدل‌سازی این گونه داده‌ها با توجه به بیش‌پراکنشی آن‌ها نسبت به توزیع پواسن و داشتن وابستگی سریالی مثبت قوی نامناسب است (۱۲،۱۴،۱۵). با توجه به این‌که در بیش‌تر داده‌های شمارشی ممکن است برخی دوره‌های زمانی دارای نرخ کم رخداد و برخی دوره‌ها دارای نرخ نسبتاً بالایی باشند، یعنی تعداد رخدادها از یک توزیع نیامده باشند بلکه آمیخته‌ای از چندین توزیع باشند که در این صورت یک ابزار آماری مناسب برای مدل‌بندی این نوع داده‌های شمارشی که هر دو شرایط بیش‌پراکنشی و وابستگی سریالی داده‌ها را به حساب می‌آورد HMM است (۱۵،۱۶). این مطالعه نشان داد که یک HMM دوحالتی با توجه به معیار مناسبت مدل، بهترین برازش را به داده‌ها دارد؛ به‌طوری که تعداد موارد جدید سل در استان همدان از دو توزیع پواسن با نرخ رخداد ۱۰/۲ و ۵/۹۶ در ماه پیروی می‌کنند. اگرچه برای مجموعه داده‌های حاضر، یافته‌های برازش مدل پواسن استاندارد نشان داد که داده‌ها از توزیع پواسن با میزان وقوع موارد جدید ۹/۳ در ماه پیروی می‌کنند، اما HMM برازش این مدل را تأیید نکرد. بنابراین عدم ارزیابی فرضیه‌های مدل ممکن است منجر به انتخاب مدل نامناسب شود.

در مطالعه رافعی و همکاران (۲۰۱۲) که برای برآورد هفتگی

موارد جدید سل ریوی اسمیر مثبت در ایران با استفاده اطلاعات سال‌های ۹۰-۱۳۸۵ انجام شد، یافته‌ها نشان داد که یک مدل پواسن دوحالتی به داده‌ها برازش مناسبی دارد یعنی با استفاده از آمیخته‌ای از دو توزیع پواسن می‌توان به پیش‌بینی موارد جدید بیماران سل ریوی اسمیر مثبت در ایران پرداخت (۲۰). یافته‌های مطالعه رافعی و همکاران با این مطالعه از نظر تعداد حالت‌ها و روش به‌کار رفته هم‌سو است.

مددی‌زاده و همکاران از HMM برای پیش‌بینی حالت‌های مختلف بیماری کبدی استفاده کردند. آن‌ها از یک مدل مارکوف زمان گسسته ۵ حالتی استفاده کردند و مانند این مطالعه روش EM را یرای برآورد پارامترها به‌کار بردند (۲۱).

والیس و همکاران در سال ۲۰۰۸ میلادی از HMM برای آشکار کردن جنبه‌های پنهان مختلفی از سل، مانند زمان تا شروع (حمله) توبرکلوز در بیماران درمان شده با عامل نکروز دهنده تومور برای تعیین اثرات درمان روی فعالیت مجدد و پیشرفت TB از یک مدل ۵ حالتی استفاده کردند (۳۲).

هم‌چنین والیس و همکاران در سال ۲۰۱۶ میلادی، روش HMM را برای پیش‌بینی مخاطره عود بیماری در افراد مبتلا به سل از یک مدل مارکوف ۵ حالتی در یک مطالعه کارآزمایی بالینی به‌کار بردند (۱۲).

لی و همکاران (۲۰۱۳) در چین از HMM برای پیش‌بینی روند پیشرفت بیماری در مبتلایان به سرطان ریه استفاده کردند. ساختار داده‌های آن‌ها یک مدل مارکوف سه حالتی را پیشنهاد داد. با استفاده از این مدل روند پیشرفت بیماری برای بیماران جدید با احتمالات به‌دست آمده از روی ماتریس احتمال انتقال قابل پیش‌بینی بود. یافته مطالعه آن‌ها مزیت HMM را به سایر مدل‌ها در پیش‌گویی حالت‌های آتی بیماری نشان داد که می‌تواند به پزشکان در تشخیص به‌موقع بیماری پیش از ورود به مرحله بعدی کمک کند (۱۹).

هم‌چنین مدل HMM در تحلیل سیگنال الکتروکاردیوگرافی، تحلیل فراوانی موارد صرع، تحلیل دنباله DNA و مدل‌بندی داده‌های نرخ بیماری شبه آنفلوانزا و فلج اطفال به‌کار برده‌اند (۳۳-۳۶).

دبان و همکاران از مدل مارکوف چندمتغیره برای پیش‌بینی تعداد موارد جدید TB در ایالات مختلف آمریکا استفاده کردند. آن‌ها از اطلاعات ۲۰۰۰-۱۹۸۰ میلادی برای پیش‌بینی موارد جدید TB تا سال ۲۰۱۰ میلادی استفاده کردند. یافته‌های مطالعه آن‌ها نشان داد که بر اساس مدل ارایه شده، میزان بروز در تمام ایالات و زیرگروه‌های جمعیتی دارای روندی نزولی است و میزان بروز از ۱۸-۵ نفر در هر صدهزار نفر در ایالت‌های مختلف در سال ۲۰۱۰ میلادی پیش‌بینی شد (۳۷). در این مطالعه از مدل مارکوف استفاده شده و فرض شده که موارد جدید بیماری از یک توزیع (یک جامعه) آمده‌اند، اما در این مطالعه براساس مدل HMM مشخص شد که داده‌ها از دو توزیع با پارامترهای مختلف هستند.

در مطالعه‌ای که توسط کلوتز (۲۰۱۳) در ایالت کبک کانادا در پیش‌بینی تعداد موارد سل در سه زیرجامعه مهاجران، سرخ پوست‌ها و غیربومی متولد شده در کانادا با استفاده از معادله‌های تفاضلی بر اساس اطلاعات ۲۰۰۹-۱۹۸۵ میلادی انجام شد یافته های مطالعه نشان داد که تعداد موارد جدید طی سال‌های آتی برای سال‌های ۲۰۳۰-۲۰۱۵ میلادی برای متولدهای غیربومی تقریباً به صفر رسیده و در مهاجران کاهشی و در بین سرخپوستان افزایش نسبی داشته است، اما در کل جمعیت ایالت کبک، تعداد موارد جدید دارای روند کاهشی بوده و تعداد موارد جدید در سال ۲۰۳۰ کم‌تر از ۳۰ مورد پیش‌بینی شد (۲۲).

وودروف و همکاران (۲۰۱۳) در مطالعه‌ای در آمریکا از اطلاعات تعداد مبتلایان به سل طی سال‌های ۲۰۱۰-۲۰۰۰ میلادی برای پیش‌بینی تعداد مبتلایان در سال‌های ۲۰۲۰-۲۰۱۱ میلادی از مدل رگرسیون لجستیک استفاده کردند. در برازش مدل، جمعیت مبتلایان به دو زیرجامعه متولدهای آمریکا و متولدهای خارج از کشور تقسیم شدند. یافته‌های پیش‌بینی نشان از روند کاهشی در تعداد موارد جدید را نشان داد که این روند کاهشی در متولدهای آمریکا نسبت به افراد مبتلا که در خارج از آمریکا متولد شده بودند، دارای شیب بسیار شدیدتری بود و روند کاهشی در متولدهای آمریکا شیب بسیار کندی داشته است (۲۳).

ژانگ و همکاران (۲۰۱۳) در پیش‌بینی رخداد سل بر اساس اطلاعات استان هوبی چین که دارای بار بیماری بالایی بود از مدل ARIMA فصلی و ARIMA رگرسیون شبکه عصبی تعمیم یافته استفاده کردند. آن‌ها از اطلاعات موارد جدید این بیماری مربوط به سال‌های ۲۰۱۱-۲۰۰۴ میلادی برای پیش‌بینی سال ۲۰۱۲ استفاده کردند. یافته‌های ارزیابی مناسبت مدل، کارایی بهتر مدل ARIMA رگرسیون شبکه عصبی تعمیم یافته را نسبت به مدل ARIMA فصلی نشان داد. مدل برازش یافته نشان داد که که تعداد موارد جدید در سال ۲۰۱۲ نزولی است (۲۴). هم‌چنین کائو و همکاران (۲۰۱۳) از مدل سری زمانی ARIMA فصلی و مدل ARIMA– شبکه عصبی مصنوعی برای پیش‌بینی تعداد موارد جدید TB از اطلاعات ۲۰۱۰-۲۰۰۵ میلادی در چین استفاده

کردند. یافته‌های شبیه‌سازی نشان داد که مدل ARIMA – شبکه عصبی مصنوعی عملکرد بهتری دارد. مدل ارایه شده نشان داد که مدل تعداد موارد جدید سالانه TB در چین روند اندکی کاهشی دارد، اما روند ماهانه دارای تغییرات دوره‌ای در برخی از مناطق این کشور بوده و بیش‌ترین مورد جدید معمولاً در فصل بهار که جشن‌های بزرگ در این کشور وجود دارد و شهرهایی که مسافران بیش‌تری را پذیرا هستند، مشاهده می‌شود (۲۵). در مطالعه دیگری که توسط چن و همکاران در برخی از ایالت‌های چین انجام دادند از مدل ARIMA برای پیش‌بینی تعداد موارد جدید سل ریوی براساس اطلاعات ۲۰۱۰-۲۰۰۴ میلادی استفاده کردند. یافته مطالعه آن‌ها نشان داد که در سال ۲۰۱۱ میلادی یک روند نزولی در میزان بروز موارد جدید سل ریوی وجود دارد (۲۶). در داده‌های این مطالعه هم طی سال‌های۹۴-۱۳۸۴ روند فصلی در موارد جدید مشاهده شد که بیش‌ترین موارد جدید مربوط به بهمن و اسفند و فصل بهار بوده است.

شکری و همکاران (۲۰۱۴) به پیش‌بینی موارد جدید TB در عربستان سعودی با استفاده از مدل رگرسیونی خطی تعمیم یافته شامل مدل رگرسیون پواسن و دوجمله‌ای منفی پرداختند. براساس تعداد موارد جدید TB طی سال‌های ۲۰۰۹-۲۰۰۰ میلادی آن‌ها مبادرت به پیش‌بینی تعداد موارد جدید برای سال‌های ۲۰-۲۰۱۰ میلادی نمودند. تعداد موارد جدید طی سال‌های پیش‌بینی دارای روند افزایشی بوده و این روند در شهرهای مهاجرپذیر بیش‌تر بوده و علت این افزایش را ورود کارگران مهاجر ذکر کردند (۲۷). با توجه به این‌که افرادی که با استان همدان وارد می‌شوند بیش‌تر از غرب کشور (استان‌های ایلام، کردستان و کرمانشاه) هستند و آمار ابتلا در این استان‌ها تفاوت زیادی با استان همدان ندارد و با توجه به این‌که بالاترین آمار مبتلایان مربوط به استان‌های سیستان و بلوچستان و گلستان بوده و از این استان‌ها مهاجر زیادی در استان همدان وجود ندارد؛ بنابراین ثابت بودن تعداد موارد جدید پیش‌بینی شده منطقی به نظر می‌رسد.

موسی‌زاده و همکاران (۲۰۱۴) از مدل ARIMA فصلی باکس-جنکینز و از اطلاعات سال‌های ۲۰۱۱-۲۰۰۵ میلادی برای پیش‌بینی موارد جدید TB استفاده کردند. نتیجه مطالعه آن‌ها روند افزایشی برای موارد جدید مبتلایان به TB را نشان داد؛ به‌طوری که میزان بروز موارد جدید براساس مدل ارایه شده از ۱۳/۷۸ در صدهزار در سال ۲۰۱۱ میلادی به ۱۶/۷۵ در سال ۲۰۱۴ نفر پیش‌بینی شد (۲۸). همچنین موسی‌زاده و همکاران (۲۰۱۵) از اطلاعات بیماران موارد جدید TB اسمیر مثبت طی سال‌های ۱۲-۲۰۰۵ میلادی برای پیش‌بینی تا سال ۲۰۱۵ میلادی از مدل ARIMA فصلی استفاده کردند. نتیجه مطالعه آن‌ها یک روند افزایشی برای موارد جدید مبتلایان به TB اسمیر مثبت را نشان داد و مدل ارایه شده میزان بروز موارد جدید را ۹/۸ در هر صدهزار پیش‌بینی کرد (۲۹). در هر دو مطالعه موسی‌زاده، روند افزایشی برای موارد جدید هم در کل مبتلایان و هم برای مبتلایان سل ریوی اسمیر مثبت پیش‌بینی شد، اما در این مطالعه روند تعداد موارد جدید در کل مبتلایان تقریباً ثابت به‌دست آمد.

در مطالعه هولو و همکاران که روی مبتلایان به TB در ۲۹ کشور اتحادیه اروپا با هدف پیش‌بینی تعداد موارد جدید در سال‌های ۲۵-۲۰۱۶ میلادی انجام شد، از مدل رگرسیون لگ- خطی و با فرض نرخ تغییرات ثابت استفاده شد. پیش‌بینی آن‌ها به تفکیک دو زیرجامعه (بیمارانی که در اروپا متولد شده و بیمارانی که خارج از اروپا متولد شده بودند) انجام شد. یافته‌های مطالعه آن‌ها نشان داد که کاهش در میزان بروز موارد جدید در متولدین اروپا ۷ درصد و در متولدهای خارج از اروپا ۳/۷ درصد بوده است، اما در کل مبتلایان ۵/۳ درصد کاهش در میزان بروز در کل جمعیت اتحادیه اروپا مشاهده شد. هم‌چنین براساس مدل ارایه شده، میزان بروز دارای روند کاهشی بوده و این میزان در متولدین اروپا ۴/۳ نفر در صدهزار و در گروه دیگر در حدود ۲۰ در هزار بوده است (۳۰). در این مطالعه با توجه به این‌که آمار موارد جدید به تفکیک بومی-غیربومی وجود نداشت، بنابراین امکان پیش‌بینی روند تغییرات به تفکیک بومی-غیربومی وجود نداشت تا بتوان با سایر مطالعه‌ها مقایسه نمود.

از جمله محدودیت‌های این مطالعه این بود که تنها فراوانی موارد جدید ماهانه سل برای دوره زمانی ۱۱ سال موجود بود که داشتن اطلاعات برای دوره‌های زمانی طولانی‌تر می‌تواند به برآورد حالت‌های ایستاتر مفید باشد. هم‌چنین در صورت وجود عوامل مربوط به بیماری و عوامل فردی در مجموعه داده می‌توان از مدل‌های مارکوف پنهان با کوواریت نیز برای بهبود برازش و پیش‌بینی مدل استفاده کرد.

نتیجه‌گیری

این مطالعه یک HMM دوحالتی را برای پیش‌بینی موارد جدید سل ارایه داد که نشان می‌دهد داده‌های موجود آمیخته‌ای از دو توزیع پواسن است. در صورتی که داده‌ها دارای پیش‌فرض‌های لازم (نرمال بودن، بیش‌پراکنشی و ...) نباشد روش‌های تحلیل

سری زمانی و مدل‌های رگرسیونی قابل استفاده نیستند و مدل HMM می‌تواند جایگزین مناسبی در پیش‌بینی داده‌های سلامت باشد که قادر است با استفاده از زنجیر مارکوف آمیختگی موجود در داده‌ها را که مدل‌های معمول قادر به تشخیص آن نیستند شناسایی کند. هم‌چنین استفاده از این مدل باعث ارایه ماتریس احتمال انتقال بین حالت‌های مختلف بیماری است که به پزشکان در پیش‌بینی مراحل آتی بیماری‌ها کمک می‌کند تا پیش از ورود به آن مرحله اقدامات لازم را در راستای پیش‌گیری و درمان بهتر بیماری انجام دهند.

تشکر و قدردانی

بدین وسیله از مسؤولان محترم معاونت بهداشتی دانشگاه علوم پزشکی همدان که داده‌های این مطالعه را در اختیار پژوهشگران قرار دادند تشکر و قدردانی می‌شود.

منابع

1- Steingart KR, Henry M, Ng V, Hopewell PC, Ramsay A, Cunningham J, et al. Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. Lancet Infect Dis 2006; 6: 570-81.

2. Global tuberculosis control : epidemiology, strategy, financing : WHO report 2009; 6-34.

3. Nasehi M, Mirhaghani L. National guidelines for TB control. Iran Ministry of Health and Medical Education; 2009; 19-20.

4. Velayati AA, Masjedi MR, Farnia P, Tabarsi P, Ghanavi J, ZiaZarifi A, et al. Emergence of new forms of totally drug-resistant tuberculosis bacilli. Chest. 2009; 136: 420-5.

5. Velayati AA, Farnia P, Masjedi MR, Ibrahim T, Tabarsi P, Haroun R, et al. Totally drug-resistant tuberculosis strains: evidence of adaptation at the cellular level. ERJ. 2009; 34: 1202-3.

6. Liu L, Zhao X, Zhou Y. A tuberculosis model with seasonality. Bull Math Biol. 2010; 72: 931-52.

7. Rios M, Garcia J, Sanchez J, Perez D. A statistical analysis of the seasonality in pulmonary tuberculosis. Eur J Epidemiol. 2000; 16: 483–8.

8. Shmueli G, Burkom H. Statistical challenges facing early outbreak detection in biosurveillance. Technometrics. 2010; 52: 39–51.

9. Castillo-Chavez C. Infectious Disease Informatics and Biosurveillance2010.

10. Unkel S, Farrington C, Garthwaite P, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. J Roy Statist Soc Ser A. 2011; 175: 49-82.

11. Lu H, Zeng D, Chen H. Prospective infectious disease outbreak detection using Markov switching models. IEEE T Knowl Data En. 2009; 22: 565-77.

12. Wallis R. Mathematical Models of Tuberculosis Reactivation and Relapse. Front Microbiol. 2016; 7: 1-7.

13. Watkins R, Eagleson S, Veenendaal B, Wright G, Plant A. Disease surveillance using a hidden Markov model. BMC Med Inform Decis Mak. 2009; 9: 39-45.

14. Zucchini W, MacDonald I. Hidden Markov models for time series: an introduction using R: Chapman & Hall/CRC; 2009.

15. Altman RM, Petkau JA. Application of hidden Markov models to multiple sclerosis lesion count data. Statist Med. 2005; 24: 2335-44.

16. Lu Y, Zeng L. A nonhomogeneous Poisson hidden Markov model for claim counts. ASTIN Bulletin 2012; 42: 181-202.

17. Le Strat Y, Carrat F. Monitoring epidemiologic surveillance data using hidden Markov models. Stat Med. 1999; 18: 3463-78.

18. Vimala K. Stress causing Arrhythmia Detection from ECG Signal using HMM. IJIRCCE; 2014. 2: 6079-85.

19. Li HM ,Fang LY, Wang P, Yan JZ. Hidden Markvo Models Based Research on Lung Cancer Progress Modeling. Research Journal of Applied Sciences, Engineering and Technology; 2013; 6: 2470-73.

20. Rafei A, Pasha E, Jamshidi Orak R. Tuberculosis Surveillance Using a Hidden Markov Model. Iran J Public Health. 2012; 41: 87–96.

21. Madadizadeh F, Montazeri M, Bahrampour A. Predicting of liver disease using Hidden Markov Model, Razi Journal of Medical Sciences,2016, 23: 66-74.

22. Klotz A, Harouna A, Smith AF. Forecast analysis of the incidence of tuberculosis in the province of Quebec. BMC Public Health 2013, 13: 400.

23. Woodruff RSY, Winston CA, Miramontes R, Predicting U.S. Tuberculosis Case Counts through 2020. PLoS ONE. 2013; 8: e65276. doi:10.1371/Journal.pone.0065276

24. Zhang G, Huang S, Duan Q, Shu W, Hou Y, et al. Application of a Hybrid Model for Predicting the Incidence of Tuberculosis in Hubei, China. PLoS ONE. 2013; 8: e80969. doi:10.1371/journal.pone.0080969

25. Cao S, Wang F, Tam W, Tse LA, Kim JH, Liu J, Lu Z: A hybrid seasonal prediction model for tuberculosis incidence in China. BMC Medical Informatics and Decision Making. 2013; 13: 56. 2-7.

26. Chen YP, Wu AP, Wang CL, Zhou HY, Feng SX, Time Series Analysis of Pulmonary Tuberculosis Incidence: Forecasting by Applying the Time Series Model, Advanced Materials Research, 2013; 709: 819-22.

27. Shoukri MM, Varghese B, Al-Hajoj S, Al-Mohanna F. Prediction of the Number of Tuberculosis Cases and Estimation of Its Treatment Cost in Saudi Arabia Using Proxy Information. Open Journal of Statistics,2014; 4, 726-35. http://dx.doi.org/10.4236/ojs.2014.49067

28. Moosazadeh M, Nasehi M, Bahrampour A, Khanjani N, Sharafi S, et al. Forecasting Tuberculosis Incidence in Iran Using Box-Jenkins Models, Iran Red Crescent Med J. 2014; 16: e11779. doi: 10.5812/ircmj.11779.

29. Moosazadeh M, Khanjani N, Nasehi M, Bahrampour A. Predicting the Incidence of Smear Positive Tuberculosis Cases in Iran Using Time Series Analysis, Iran J Public Health. 2015; 44: 1526-34.

30. Hollo V, Beaute J, Kodmon C, van der Werf M. Tuberculosis notification rate decreases faster in residents of native origin than in residents of foreign origin in the EU/EEA, 2010 to 2015. Euro Surveill. 2017; 22: pii=30486. DOI: http://dx.doi.org/10.2807/1560-7917.ES.2017.22.12.30486

31. Le Strat Y, Carrat F. Monitoring epidemiologic surveillance data using hidden Markov models. Statistics in medicine. 1999; 18: 3463-78.
32. Wallis RS. Mathematical modeling of the cause of tuberculosis during tumor necrosis factor blockade. Arthritis Rheum. 2008; 58: 947-52.
33. Rath T, Carreras M, Sebastiani P. Automated detection of influenza epidemics with hidden Markov models. Chest. 2003: 521-32.
34. Jamshidi Orak R, Mohammad K, Pasha E, Sun W, Nori Jalyani K, Rasolinejad M, et al. Modeling the spread of infectious diseases based the Bayesian approach. Journal of School of Public Health and Institute of Public Health Research. 2007; 5: 7-15.
35. McBryde E, Pettitt A, Cooper B, McElwain D. Characterizing an outbreak of vancomycin-resistant enterococci using hidden Markov models. Journal of The Royal Society Interface. 2007; 4: 745-54.
36. Held L, Hofmann M, MH, Schmid V. A two-component model for counts of infectious diseases. Biostatistics. 2006; 7: 422-37.
37 .Debanne SM., Bielefeld RA, Cauthen GM, Danieln TM, Rowland DY. Multivariate Markovian Modeling of Tuberculosis: Forecast for the United States, Emerging Infectious Diseases. 2000, 6: 148-57.

Iranian Journal of Epidemiology Summer 2018; 14 (2): 126-135.
Original Article

Application of Hidden Markov Model in Forecasting New Cases of Tuberculosis in Hamadan Province Based on the Recorded Cases during 2006-2016

Safari M[1], Sadeghifar M[2], Roshanaei Gh[3], Zahiri A4[4]

1-PhD Candidate, Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran
2- Assistant Professor in Statistics, Department of Mathematics, Bu-Ali-Sina University, Hamadan, Iran
3- Associate Professor in Biostatistics, Modeling of Noncommunicable Disease Research Canter, Hamadan University of Medical Sciences, Hamadan, Iran
4- BSc of Public Health Center for Disease Control & Prevention, Deputy of Health Services, Hamadan University of Medical Sciences, Hamadan, Iran
Corresponding author: Roshanaei GH, gh.roshanaei@umsha.ac.ir
(Received 25 November 2017; **Accepted** 21 Aprilr 2018)

Background and Objectives: Tuberculosis is a chronic bacterial disease and a major cause of morbidity and mortality. It is caused by a Mycobacterium tuberculosis. Awareness of the incidence and number of new cases of the disease is valuable information for revising the implemented programs and development indicators. time series and regression are commonly used models for prediction but these methods require some assumptions. The purpose of this study was to predict new TB cases using the hidden Markov model which does not require many assumption.

Methods: The data used in this study was the monthly number of new TB cases during 2006-2016 identified and recorded in Hamedan Province. Rorecasting the number of new TB cases was done using hidden Markov models using the hidden Markov package in the R software.

Results: According to the AIC and BIC criterion, two states had the best fit to the data, i.e. the data of this study were a mixture of two Poisson distributions with average number of event 5.96 and 10.2 respectively. The results also predicted the number of new cases over the next 24 months based on the hidden Markov model would be between 8 and 9 new cases in each month.

Conclusion: The hidden Markov model is the best model for prediction using the Markov chain. This model, in addition to detection of an appropriate model for the available data, can determine the transition probability matrix, which can help physicians predict the future state of the disease and take preventive measures befor reaching advanced stages.

Keywords: Tuberculosis, Hidden Markov model, Prediction, Hamadan