УДК 681.3

О.Н. Чопоров, О.В. Золотухин, С.В. Болгов

АЛГОРИТМИЗАЦИЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ О РАСПРОСТРАНЕННОСТИ ЗАБОЛЕВАНИЙ НА РЕГИОНАЛЬНОМ И МУНИЦИПАЛЬНОМ УРОВНЯХ

Воронежский институт высоких технологий
Воронежский государственный медицинский университет им. Н.Н. Бурденко
Орловский государственный университет

В рамках многоуровневого мониторинга распространенности заболеваний предлагается последовательность этапов выполнения интеллектуального анализа данных: 1) формирование информационной базы данных; 2) сравнительный анализ данных по отдельным территориальным единицам; 3) анализ текущего состояния, динамики исследуемых показателей и прогнозирование их изменения; 4) анализ взаимосвязи показателей заболеваемости населения, деятельности и ресурсного обеспечения психиатрической службы; 5) классификация районов области по уровню заболеваемости населения исследуемой патологией. Для каждого этапа предложен комплекс методов и средств, позволяющих выполнить поставленные задачи.

Ключевые слова: анализ заболеваемости, многоуровневый мониторинг, анализ временных рядов, экспоненциальное сглаживание, классификация территориальных единиц.

Одной из основных предпосылок, обеспечивающих рациональное планирование и управление системой здравоохранения, является организация интерактивного сбора, поиска, накопления разнородной информации, а также предоставление возможности получения наглядной информации, характеризующей состояние здоровья населения, в реальном масштабе времени, что достигается посредством применения систем мониторинга.

При этом мониторинг предлагается рассматривать как многоуровневую систему с выделением регионального (муниципального) и индивидуального уровней. На региональном (муниципальном) уровнях анализируются показатели заболеваемости населения, деятельности и ресурсного обеспечения учреждений здравоохранения, а на индивидуальном – медико-социальные факторы риска и состояние здоровья.

При использовании многоуровневого подхода для системного анализа результатов мониторинга на каждом уровне рассмотрения требуется выбор адекватных методов статистической обработки данных, математического моделирования и принятия решений [7].

На региональном уровне мониторинг заболеваемости и показателей, характеризующих медицинское обслуживание населения региона, а также

отдельных его территориальных единиц (или районов области) требуется не только для оценки сложившейся ситуации, но и для построения краткосрочных прогнозов. Результаты мониторинга являются основой для формирования групп районов с неблагоприятной и благоприятной ситуацией, и могут быть использованы при принятии адекватных управленческих решений на соответствующем уровне.

К показателям заболеваемости населения относятся

- общая заболеваемость,
- первичная заболеваемость,
- число больных,
- контингенты больных,
- число больных, состоящих на диспансерном учете и др.

К показателям деятельности учреждений здравоохранения относятся

- работа койки соответствующего профиля,
- оборот койки,
- средняя длительность пребывания больного в стационаре,
- летальность и др.

К показателям ресурсного обеспечения относятся

- объемы финансирования (из бюджета, по ОМС, от платных услуг),
- обеспеченность врачами,
- обеспеченность средним и младшим мед. персоналом,
- обеспеченность койками,
- укомплектованность врачебными кадрами и др.

Оценка текущего состояния уровня заболеваемости населения, деятельности, ресурсного обеспечения учреждений здравоохранения или прогноза их изменения во времени возможны на основе анализа перечисленных характеристик, однако при этом возможно только сравнение по отдельным показателям, а комплексная оценка весьма затруднена. Подобная оценка требуется при выборе среди альтернативных управленческих решений, когда имеется прогноз изменения отдельных анализируемых показателей, но они не позволяют охарактеризовать сложившуюся ситуацию в целом. В связи с этим требуется разработка интегральных показателей, позволяющих комплексно оценить сложившуюся ситуацию с учетом отдельных показателей и их значимости. Интегральные показатели могут быть разработаны отдельно для описания заболеваемости населения, деятельности и ресурсного обеспечения учреждений здравоохранения.

Для расчета интегрального показателя предлагается следующая формула:

$$\text{ИП} = \sum_{\text{i=1}}^{\text{N}} w_i X_i^{\text{н}} \qquad (1)$$

где N – количество показателей, включенных в интегральный;

w_i – значимость (вес) i-го показателя,

$X_i^{\text{н}}$ – нормированное значение (балльная оценка) i-го показателя.

Для определения значимости каждого показателя, включенного в интегральный, используется метод априорного ранжирования [2, 4].

Значения весов wi определяются по формуле

$$w_i = \frac{m \cdot (n+1) - \sum_{j=1}^{m} r_{ij}}{\sum_{i=1}^{n} \left(m \cdot (n+1) - \sum_{j=1}^{m} r_{ij} \right)}, \; i = \overline{1,n}. \qquad (2)$$

где r_{ij} $(j = \overline{1,m})$ – ранг, выставленный j-м экспертом.

С учетом того, что $\sum_{i=1}^{n} w_i = 1$, а показатели, включенные в ИП, приведены к нормированному виду или оценены с использованием k-балльной шкала, максимальное значение интегрального показателя соответствует верхней границе нормировки (обычно равной +1) или максимальному баллу (k), а минимальное значение равно нижней границе нормировки (обычно это 0) или минимальному баллу (обычно +1).

Для оценки ситуации, сложившейся за анализируемый период рассчитываются средние значения показателей. Полученные результаты анализа заболеваемости населения, показателей медицинского обслуживания и разработанные на их базе интегральные показатели, дают возможность получить оценку текущего состояния ситуации в исследуемых территориальных единицах региона (районах области). Однако, существенный интерес представляет собой оценка динамики анализируемых показателей, а также результаты краткосрочного и долгосрочного прогнозирования их изменения.

При анализе динамики по каждому анализируемому показателю рассчитывается цепной темп прироста $T_{пр}^{Ц}$ и базисный прирост $T_{пр}^{Б}$ относительно начала наблюдения. Для расчетов используются следующие выражения:

$$T_{пр}^{Ц}(t) = \frac{y_t - y_{t-1}}{y_{t-1}} \cdot 100\%, \quad \text{t=}\overline{\text{н+1,к}} \qquad (3)$$

$$T_{пр}^{Б} = \frac{y_к - y_н}{y_н} \cdot 100\%, \quad (4)$$

где *t* – порядковый номер измерения показателя во временном интервале; *н* – год начала исследования; *к* – год окончания исследования.

Помимо этого при оценке динамики показателей рассчитывается цепной абсолютный прирост ($\Delta^{Ц}$) и базисный абсолютный прирост ($\Delta^{Б}$):

$$\Delta^{Ц} y_t = y_t - y_{t-1}, \quad t = \overline{н+1, к} \quad (5)$$

$$\Delta^{Б} y = y_к - y_н \quad (6)$$

Для обеспечения проведения сравнительного анализа и определения районов с неблагоприятной и благоприятной ситуацией, данные, представленные на районном уровне, должны быть приведены к стандартизованным единицам измерения. В нашем исследовании предложено проводить нормировку анализируемых показателей относительно их среднего значения.

При исследовании причинно-следственных связей, имеющихся между состоянием здоровья населения, ресурсным обеспечением и деятельностью медицинских учреждений и выработки вариантов управленческих решений, необходимо проведение анализа статистических взаимосвязей (корреляционного анализа) между выделенными группами показателей.

Принимаемые управленческие решения зависят как от результатов анализа сложившейся ситуации, так и от результатов краткосрочного прогнозирования изменения исследуемых показателей.

Краткосрочный прогноз – это прогноз «на завтра», то есть, прогноз на несколько шагов вперед. При более формальном подходе можно сказать, что под данным понятием подразумевается построение прогноза на 1-3 шага вперед. Для данного типа прогноза предлагается использовать модель временного ряда, построенную с использованием метода экспоненциального сглаживания, который является одним из самых распространенных и эффективных способов, используемых при исследовании и прогнозировании в медицинских исследованиях. Достоинство данного метода это его адаптивность – с его помощью учитывается информация, содержащаяся в уровнях временного ряда, дифференцированно, то есть, чем информация «старше», тем меньше ее ценность для прогнозирования.

Для временного ряда может быть использовано представление в виде аддитивной модели

$$y_t = m + \varepsilon_t, \quad (7)$$

где y_t – значение анализируемого показателя в момент времени t; $m = \overline{y_t}$ - среднее значение (математическое ожидание); ε_t – некоторая случайная составляющая.

При сглаживании уровней ряда y_t применяется рекуррентная формула

$$S_t = \alpha y_t + (1-\alpha) S_{t-1}, \quad t = \overline{1,\mathrm{N}}\ , \tag{8}$$

где S_t – сглаженное значение реального уровня y_t в момент времени t; α - коэффициент сглаживания ($0<\alpha<1$); N – число измерений (длина ряда).

Параметр сглаживания α выбирается исследователем и представляет собой весовой коэффициент для фактического уровня y_t. Если ряд стабильный, с малым значением случайной составляющей ε_t, целесообразно выбирать значения α, близкие к 1 ($S_t \approx y_t$). Если имеется существенное влияние ε_t на значение y_t, целесообразно выбирать α, близкие к 0, так как в этом при этом выполняется максимальное сглаживание уровней ($S_t \approx S_{t-1}$).

Обычно значение S_0 приравнивается первому из уровней: $S_0 = y_1$. Для повышения качества прогноза для вычисления S_0 предлагается использовать следующее выражение:

$$S_0 = y_1 - \frac{T_0}{2}, \tag{9}$$

где $T_0 = \left(y_N - y_1\right) \big/ (N-1)$.

Вычисление будущего прогнозного значения y_{t+1} выполняется на основе следующего выражения:

$$y_{t+1} = S_t. \tag{10}$$

Однако, для временного ряда, который имеет выраженный тренд, что часто бывает с большинством медико-социальных показателей, обычное экспоненциальное сглаживание малоэффективно, так как игнорирование трендовой зависимости приводит к накоплению систематической ошибки в результатах. Поэтому в исходной модели простого экспоненциального сглаживания требуется коррекция на тренд. В связи с этим, при построении прогностических моделей предлагается использовать одну из наиболее эффективных моделей экспоненциального сглаживания с учетом тренда – *модель Хольта* [3]. В данной модели тренд тоже подвергается процедуре сглаживания. Процедура сглаживания тренда основана на рекуррентном соотношений, аналогичном (8), но с другим персональным параметром γ:

$$\begin{cases} S_t = \alpha y_t + (1-\alpha)(S_{t-1} + b_{t-1}); \\ b_t = \gamma(S_t - S_{t-1}) + (1-\gamma) b_{t-1}, \end{cases} \quad (11)$$

где t – временные отсчеты, $t = 1, 2, \ldots$;

α, γ – параметры сглаживания ($0<\alpha<1$; $0<\gamma<1$);

y_t – реальный уровень ряда в момент времени t;

S_t – сглаженное значение реального уровня y_t;

b_t – сглаженное значение тренда для момента времени t.

Для запуска вычислительного процесса в рекуррентных формулах (11) требуется задать начальные значения S_0 и b_0, а также выбрать наиболее подходящие параметры для сглаживания α и γ. На начальном этапе значение тренда b_0 обычно выбирается равным нулю. В предлагаемой модели Хольта прогнозируемые значения определяются на основе формулы

$$y_{t+1} = S_t + b_t. \quad (12)$$

Существенным преимуществом методов, которые основаны на модели экспоненциального сглаживания, является наличие возможности учета временной ценности информации и адаптации к изменяющимся условиям, что имеет существенное значение при нестабильно протекающих процессах.

Следующей задачей интеллектуального анализа данных о распространенности заболеваний на региональном и муниципальном уровнях является классификация отдельных территориальных единиц (районов области) по уровню анализируемых показателей. Причем классификацию предлагается выполнять как по среднему уровню показателей за анализируемый период, так и по базисным темпам прироста [1, 5, 6].

С точки зрения сложившейся ситуации предлагается выделять 3 группы районов: 1) с высоким уровнем анализируемых показателей; 2) со среднем уровнем показателей; 3) с низким уровнем показателей.

В качестве критерия для определения границы между группами предложено использовать среднеквадратическое отклонение $\sigma(x_i)$ анализируемых показателей x_i:

- группа 1: $x_i > \sigma(x_i)$;
- группа 2: $-\sigma(x_i) \le x_i \le \sigma(x_i)$; (13)
- группа 3: $x_i < -\sigma(x_i)$.

На основе проведения такой классификации появляется возможность выделить районы с высоким и низким уровнем показателей, что является

основой для принятия эффективных управленческих решений, которые направлены на профилактику заболеваемости и снижение ее уровня.

Таким образом, можно выделить следующие этапы анализа данных о распространенности заболеваний на региональном и муниципальном уровнях: 1) формирование информационной базы данных; 2) сравнительный анализ данных по отдельным территориальным единицам; 3) анализ текущего состояния, динамики исследуемых показателей и прогнозирование их изменения; 4) анализ взаимосвязи показателей заболеваемости населения, деятельности и ресурсного обеспечения психиатрической службы; 5) классификация районов области по уровню заболеваемости населения исследуемой патологией. Перечисленные этапы и используемые для их реализации методы и средства представлены в табл. 1.

Таблица 1

Этапы и методы интеллектуального анализа данных о распространенности заболеваний на региональном и муниципальном уровнях

№ п/п	Название этапа	Используемые методы и средства
1	Формирование информационной базы данных	Анализ отчетно-учетной документации Формирование интегральных показателей Создание базы данных с использованием СУБД
2	Сравнительный анализ данных по отдельным территориальным единицам	Расчет нормированных показателей Методы математической статистики
3	Анализ текущего состояния, динамики исследуемых показателей и прогнозирование их изменения	Исследование временных рядов Построение трендов с использованием модели экспоненциального сглаживания
4	Анализ взаимосвязи показателей заболеваемости населения, деятельности и ресурсного обеспечения психиатрической службы	Корреляционный анализ Регрессионный анализ
5	Классификация районов области по уровню заболеваемости населения	Классификационный анализ

ЛИТЕРАТУРА

1. Классификация территориальных единиц региона по уровню заболеваемости взрослого женского населения миомой матки и эндометриозом на основе геоинформационных технологий / Е.Н. Коровин, Н.Н. Кудинова, М.В. Фролов, О.Н. Чопоров // Information Technology Applications. – 2013. - № 4. – С. 74-81.
2. Львович Я.Е. Моделирование биотехнических и медицинских систем / Я.Е. Львович, М.В. Фролов // Под ред. В.Н. Фролова: учеб. пособие. – Воронеж: Изд-во ВГТУ, 1994.
3. Медик В.А. Математическая статистика в медицине / В.А. Медик, М.С. Токмачев// Учеб. пособие. – М.: Финансы и статистика, 2007. – 800 с.
4. Чопоров О.Н. Оптимизация функционирования медицинских систем на основе интегральных оценок и классификационно-прогностического моделирования: дис. … д-ра техн. наук. – Воронеж, 2001. – 329 с.
5. Choporov O. Technique of information database formation for carrying out multilevel monitoring and classificatory-and-forecasting modeling / O. Choporov, A. Kurotova, I. Manakin // Information Technology Applications. – 2015. - № 1. – С. 111-123.
6. Choporov O.N. Classification and prognostic modeling in medical and social research / O.N. Choporov, N.V. Naumov, N.N. Kudinova, A.I. Agarkov // Modern informatization problems in economics and safety: Proceedings of the XVIII-th International Open Science Conference (Lorman, MS, USA, January 2013). – p. 90-93.
7. Choporov O.N. Infobase formation technology for medical systems analysis and modeling / O.N. Choporov, S.V. Bolgov, L.A.Kutashova, E.Y.Konovalova // Modern informatization problems in economics and safety: Proceedings of the XVIII-th International Open Science Conference (Lorman, MS, USA, January 2013). – p. 157-162.

O.N. Choporov, O.V. Zlotukhin, S.V. Bolgov

ALGORITHMIZATION OF DISEASE MORBIDITY DATA MINING AT REGIONAL AND MUNICIPAL LEVELS

Voronezh Institute of High Technologies
Voronezh State Medical University of N.N. Burdenko
Orlov State University

In the course of multilevel monitoring of disease morbidity the authors propose a succession of stages for carrying out data mining: 1) formation of informative database; 2) comparative data analysis according to distinct geographical units; 3) analysis of a current state, dynamics of the analyzed indices and forecasting their changes; 4) interrelation analysis of the population morbidity indices and resources provision and work of a psychiatric service; 5) classification of the Region's areas according to morbidity level of the analyzed pathology. The authors offer a complex of methods and means, allowing solving the original problems.

Keywords: morbidity analysis, multilevel monitoring, time-series analysis, exponential smoothing, geographical units' classification.