

А. И. Михальский

Оценка числа ВИЧ инфицированных лиц в населении по наблюдаемой динамике числа выявленных случаев СПИД

Введение

Оценка числа лиц, инфицированных вирусом иммунодефицита человека (ВИЧ) необходима для объективного планирования и эффективного проведения мероприятий по профилактике его распространения и по борьбе с его клиническим проявлением СПИД. Эта информация важна и для построения обоснованных оценок и прогнозов экономических и демографических последствий эпидемии ВИЧ/СПИД для отдельных стран и всего мира [1]. Непосредственная регистрация факта инфицирования человека ВИЧ проводится в лабораторных условиях с применением дорогостоящих тест-комплексов. Кровь для анализа человек сдаёт либо добровольно, либо в силу действующего законодательства. В Российской Федерации обязательную проверку на ВИЧ проходят доноры крови, биологических жидкостей, блоков и тканей, все, имеющие контакт с вирусом или его носителями, заключённые, пациенты с рядом диагнозов или с подозрением на один из них, в частности, все беременные женщины, включая прерывающих беременность. Ежегодное количество тестов на антитела к ВИЧ в России охватывает 15-17 процентов всего населения [2]. Число зарегистрированных новых случаев инфицирования ВИЧ в РФ в 2002 году составило 50,340 уменьшившись по сравнению с 2001 годом, в котором оно равнялось 88,392. По данным Министерства здравоохранения России на март 2004 года в России проживало 274,197 ВИЧ инфицированных лиц [3].

Достоверность приводимых оценок нуждается в независимой проверке, поскольку априори неясно насколько эффек-

тивно работает существующая в России система учета ВИЧ инфицированных. Кроме того, до настоящего времени основной группой риска инфицирования ВИЧ в нашей стране являются наркоманы, число которых по оценкам специалистов в настоящее время превосходит 3 млн. человек. Однако, на конец 2002 г. в наркологической службе России было зарегистрировано лишь 449 тыс. человек, что, возможно, говорит о высокой доле скрытых, невыявленных лиц. Вопрос о надёжности и корректировке официальной статистики требует объективного исследования, поскольку недооценка распространённости ВИЧ в населении страны может снизить эффективность мероприятий по борьбе с этой эпидемией. В настоящее время подобная корректировка становится особенно актуальна, поскольку по мнению многих исследователей эпидемия ВИЧ в России вышла из относительно изолированных групп риска в общее население страны [4].

Математическое моделирование может помочь построить достоверные оценки числа ВИЧ инфицированных по данным о динамике эпидемии СПИД и смертности, связанной со СПИД. Считается что на уровне популяции выявляемость заболеваемости СПИД и информация о смертности от СПИД более надёжны чем непосредственная регистрация факта ВИЧ инфицирования [5]. Пересчитывая наблюдаемые случаи СПИД в возможные случаи ВИЧ инфицирования удаётся получать отсутствующую информацию об эпидемии ВИЧ и дополнять имеющиеся данные. Поскольку СПИД является отложенным во времени результатом инфицирования ВИЧ, в англоязычной литературе подобные методы оценивания получили название back calculation и активно применяются при анализе эпидемии ВИЧ/СПИД [6, 7].

Математические модели для оценки размеров эпидемии ВИЧ/СПИД

Математическая модель связи ВИЧ инфицирования и заболеваемости СПИД

Факт возникновения заболевания – вероятностный процесс, определяемый наличием возбудителя заболевания, факторами риска и состоянием организма человека, в частности, возрастом. В случае СПИД возбудителем заболевания считается вирус ВИЧ, с момента проникновения которого в организм человека начинается период разрушения иммунной системы, в результате развивается одно из заболеваний, связанных с пониженным иммунитетом и относящихся к так называемому СПИД ассоциированному комплексу. Время, необходимое для этого, измеряется годами и достигает до 15 лет. Считается, что длительность этого инкубационного периода тем короче, чем большее количество вируса попало в кровь человека. Для случаев инфицирования в молодых возрастах характерно удлинение инкубационного периода по сравнению со случаями инфицирования в более старших возрастах [8]. Существуют публикации, свидетельствующие об увеличении длительности инкубационного периода вследствие применения антиретровирусной терапии [9].

Исчерпывающей вероятностной характеристикой процесса заболеваемости является интенсивность возникновения болезни в возрасте x , определяемая как отношение вероятности здоровому человеку заболеть на протяжении малого возрастного интервала $[x, x+dx]$ к длине этого интервала, стремящейся к нулю. Именно эта величина и характеризует процесс заболеваемости в его возрастной динамике. Для целей моделирования представляет интерес не просто вероятность заболеть в возрасте x , а вероятность заболеть в возрасте x будучи инфицированным в возрасте s . Интервал $[x,s]$ является инкубационным периодом заболевания, плотность

распределения вероятности его длительности обозначим через $L(x,s)$. Точнее, $L(x,s)dxds$ – вероятность заболеть на протяжении возрастного интервала $[x, x+dx]$ будучи инфицированным на протяжении возрастного интервала $[s, s+ds]$. Пользуясь этой функцией, запишем плотность вероятности $u(t,x)$ распределения случаев заболевания по возрастам в календарном году t в виде соотношения

$$u(t,x) = \int_0^x L(x,s) \exp\left(-\int_s^x \mu_c(t-x+\tau, \tau) d\tau\right) \psi(t-x+s, s) ds, \quad (1)$$

где $\psi(t,x)$ – плотность вероятности инфицирования человека возраста x в году t , $\mu_c(t,x)$ – смертность в возрасте x в году t . Соотношение (1) будет играть ключевую роль в дальнейшем рассмотрении. Оно имеет смысл уравнения, связывающего число случаев инфицирования с числом случаев заболевания и может использоваться как для расчёта ожидаемого числа случаев заболевания по заданной динамике случаев инфицирования ("прямая" задача), так и для оценки распределения числа случаев инфицирования по наблюдаемой динамике случаев заболевания ("обратная" задача). Функция $L(x,s)$ при этом играет роль аналогичную роли матрицы при решении системы линейных уравнений в алгебре и в большой степени влияет на качество решения поставленной задачи.

Не существует универсального выражения для распределения длительности инкубационного периода. Различные исследователи строят разные описания в зависимости от доступной информации. Наиболее простым является экспоненциальное представление для $L(x,s)$ в виде

$$L(t,x) = \lambda \exp(-\lambda(t-x)), \quad x \leq t,$$

которое соответствует гипотезе о том, что интенсивность развития заболевания не зависит ни от возраста, ни от вре-

мени, то есть игнорируется динамика развития заболевания и считается, что никакого терапевтического воздействия не оказывается. Параметр λ обратно пропорционален среднему времени длительности инкубационного периода e_{lat} ($\lambda = 1/e_{\text{lat}}$) и легко вычисляется по данным частных обследований, в которых фиксируется как время заражения, так и время постановки диагноза, например при расследовании случаев заражения ВИЧ при переливании крови. Обозначив через d_j ($j=1, \dots, n$) имеющиеся данные о длительности инкубационного периода оценку для параметра λ можно записать в виде

$$\lambda = \frac{n}{\sum_{j=1}^n d_j}.$$

Более сложные выражения для распределения длительности инкубационного периода можно получить используя формулу Вейбулла

$$L(t, x) = k\nu(t - x)^{\nu-1} \exp(-k(t - x)^\nu),$$

либо рассматривая процесс развития СПИД как марковский процесс последовательных переходов в состояние СПИД ассоциированного комплекса, а из него в состояние возникновения СПИД. Время нахождения в каждом из состояний подчиняется распределению Вейбулла. Идентификация такой модели методом максимального правдоподобия по когортным данным [8] показала существенное влияние возраста заражения на распределение длительности развития СПИД ассоциированного комплекса. Распределение длительности этапа возникновения СПИД от возраста не зависело. Другие авторы рассматривают влияние на распределение инкубационного периода не только возраста, но и календарного времени [10], лечения [11, 12]. Модели различной сложности могут быть построены для описания распре-

деления длительности инкубационного периода, однако, возможности их идентификации ограничиваются имеющейся информацией. В случае наличия минимальной информации экспоненциальное представление для распределения длительности инкубационного периода представляется оправданным.

Оценка числа инфицированных как обратная задача

Соотношение (1), связывающее распределение случаев заболевания СПИД и вероятность ВИЧ инфицирования в разные годы и разных возрастах, позволяет построить оценку числа инфицированных людей в популяции. Обозначим через $\Psi(t, x)$ число людей возраста x инфицированных ВИЧ в году t , а через $U(t, x)$ число диагнозов СПИД среди людей возраста x в году t . Заменяя выражения для плотностей вероятностей этими функциями перепишем (1) в виде

$$U(t, x) = \int_0^x L(x, s) \exp\left(-\int_s^x \mu_c(t-x+\tau, \tau) d\tau\right) \Psi(t-x+s, s) ds$$

Если учитывать вероятность $\beta(t, x)$ пропуска диагноза СПИД в возрасте x в году t , то последнее соотношение необходимо скорректировать следующим образом

$$U(t, x) = \beta(t, x) \int_0^x L(x, s) \exp\left(-\int_s^x \mu_c(t-x+\tau, \tau) d\tau\right) \Psi(t-x+s, s) ds \quad (2)$$

Поскольку инфицирование ВИЧ является причиной возникновения диагноза СПИД, то вычисление числа инфицированных лиц по числу диагностированных случаев СПИД является решением обратной задачи, что само по себе является сложной проблемой. Дело в том, что наблюдаемое число диагнозов СПИД $U(t, x)$ предопределяется интегральным эффектом числа случаев заражения ВИЧ $\Psi(t, x)$ в различные годы и малые изменения в числе случаев заражения слабо

вливают на число диагнозов СПИД. Это ведёт к тому, что при решении уравнения (2) возможны очень большие искажения в решении $\Psi(t, x)$ при неизбежных малых искажениях в функции $U(t, x)$. Это свойство неустойчивости оказывается универсальным при решении многих задач и интегральных уравнений типа (2). В математике подобные задачи называют некорректно поставленными по Адамару [13], методы их решения составляют специальный раздел математической физики.

Существует несколько подходов к проблеме решения уравнения (2) применительно к оценке числа ВИЧ инфицированных по числу диагнозов СПИД [14]. Описывая процесс инфицирования стохастическим точечным процессом можно показать, что число диагнозов СПИД описывается пуассоновским процессом, математическое ожидание которого связано с числом случаев инфицирования ВИЧ интегральным уравнением свёртки аналогичным (2) [15]. Рассматривая число инфицированных в каждом году в заданных возрастных группах как параметры функции распределения этого процесса методом максимального правдоподобия строятся оценки числа ВИЧ инфицированных людей. В силу неустойчивости решения интегрального уравнения (2) полученные оценки также неустойчивы при достаточно подробном разбиении по возрастным группам. Для стабилизации решения число инфицированных лиц можно задавать в виде зависимости от времени и возраста, содержащей малое число параметров. Например в виде кусочно полиномиальных функций – сплайнов. Величины параметров также оцениваются методом максимального правдоподобия. Кроме пуассоновского для оценки числа ВИЧ инфицированных можно используют мультиномиальное правдоподобие [16], фильтр Калмана [17], однако проблемы со стабилизацией оценок сохраняются. Использование же параметрического представления не всегда оказывается эффективной.

Альтернативой параметризации является стабилизация оценок через использование при максимизации функционала правдоподобия дополнительного стабилизирующего функционала [18]. При этом оценки числа инфицированных лиц строятся путём максимизации функционала

$$\Phi_{\alpha}(\bar{\Psi}) = \text{LnL}(\bar{\Psi}) - \alpha\Omega(\bar{\Psi}), \quad (3)$$

где $\bar{\Psi}$ – вектор искомых оценок, LnL – логарифм функционала правдоподобия, α – параметр стабилизации, Ω – стабилизирующий функционал, принимающий малое значение для “гладких” оценок и большое для “нерегулярных”. Выбор стабилизирующего функционала определяет желаемые свойства оценок. Например, широко применяемый квадратичный функционал

$$\Omega(\bar{\Psi}) = \sum_{j=2}^{n-1} (\bar{\Psi}_{j-1} - 2\bar{\Psi}_j + \bar{\Psi}_{j+1})^2$$

на равномерной временной сетке соответствует норме второй производной и его использование при больших α приводит к стремлению оценок к линейной по времени зависимости. Функционал

$$\Omega(\bar{\Psi}) = \sum_{j=2}^{n-1} (\ln \bar{\Psi}_{j-1} - 2 \ln \bar{\Psi}_j + \ln \bar{\Psi}_{j+1})^2$$

при больших α приводит к построению экспоненциальной временной зависимости [15]. В [18] исходя из байесовской постановки предложен стабилизирующий функционал вида

$$\Omega(\bar{\Psi}) = \sum_{j=2}^n (\bar{\Psi}_j^{1/p} - \bar{\Psi}_{j-1}^{1/p})^2,$$

где $p \geq 1$ является свободным параметром. Этот функционал принимает нулевое значение если все компоненты

оценки $\bar{\Psi}$ равны между собой. Существует точка зрения, что выбор конкретного вида стабилизирующего функционала определяется априорными сведениями о динамике изучаемого процесса и выбор наиболее адекватной формы должен производиться путём анализа получаемых оценок исходя из их медико-биологической состоятельности.

Выбор величины параметра стабилизации α является самостоятельной задачей, примыкающей к проблеме выбора модели по эмпирическим данным [19]. При обработке биометрической информации широкое распространение получил метод cross-validation [20]. Выбор параметра стабилизации для максимизации функционала правдоподобия по этому методу заключается в максимизации по α функционала [21]

$$Cv(\alpha) = \sum_{i=1}^n \text{Ln}L_1(\bar{\Psi}_\alpha^i),$$

где $\bar{\Psi}_\alpha^i$ - вектор значений параметров, на которых при фиксированном α достигается максимальное значение стабилизированного функционала (3), в котором не присутствует i - тое наблюдение. $\text{Ln}L_1(\bar{\Psi}_\alpha^i)$ - значение логарифма правдоподобия для i - того наблюдения на векторе значений параметров $\bar{\Psi}_\alpha^i$. Метод cross-validation хорошо зарекомендовал себя при построении непараметрической регрессии, однако при максимизации правдоподобия на выборках ограниченного объёма эффективность этого метода изучена недостаточно. В следующем разделе рассматривается подход, альтернативный методу максимизации правдоподобия, и описывается конструктивный метод выбора параметра стабилизации, учитывающий ограниченность экспериментальных данных.

Оценка числа ВИЧ инфицированных по динамике случаев СПИД

Решение интегрального уравнения

В математической физике решение интегрального уравнения понимается как функция, минимизирующая функционал невязки между наблюдаемыми данными $Y = \{Y_1, \dots, Y_n\}$ и значениями функции $U = \{U_1, \dots, U_n\}$, стоящей в левой части уравнения [13],

$$\|Y - U\|^2 = \sum_{i=1}^n (Y_i - U_i)^2 .$$

Применительно к уравнению (2) функционал невязки имеет вид

$$\|Y - U(\Psi)\|^2 = \sum_{i=1}^n \left[Y_i - \beta(t_i, x_i) \int_0^{x_i} L(x_i, s) \exp\left(-\int_s^{x_i} \mu_c(t_i - x_i + \tau, \tau) d\tau\right) \Psi(t_i - x_i + s, s) ds \right]^2$$

где Y_i - число случаев СПИД, диагностированных в году t_i в возрасте x_i . Непосредственная минимизация функционала невязки приводит к сильным искажениям в решении интегрального уравнения Ψ , для стабилизации которого используется метод регуляризации, предложенный А. Н.Тихоновым в 1943 г. при рассмотрении устойчивости обратных задач математической физики. Как и при максимизации правдоподобия стабилизация решения заключается в введении дополнительного функционала Ω , принимающего малое значение для “гладких” решений и большое для “не-

регулярных". Регуляризованный функционал невязки принимает вид

$$N_{\alpha}(\Psi) = \|Y - U(\Psi)\|^2 + \alpha\Omega(\Psi). \quad (4)$$

При решении задач математической физика широко применяются стабилизирующие функционалы в виде квадратичной нормы первой производной

$$\Omega(\Psi) = \int_a^b \left(\frac{d}{d\tau} \Psi(\tau) \right)^2 d\tau,$$

второй производной

$$\Omega(\Psi) = \int_a^b \left(\frac{d^2}{d\tau^2} \Psi(\tau) \right)^2 d\tau.$$

Интересно использование статистических аналогий для стабилизации положительных решений [22]. В качестве стабилизирующего функционала принимается "энтропия" решения

$$\Omega(\Psi) = - \int_a^b \Psi(\tau) \ln \Psi(\tau) d\tau.$$

В этом случае минимизация регуляризованного функционала невязки (4) соответствует поиску среди решений с ограниченной невязкой $\|Y - U(\Psi)\|^2 \leq \delta$ решения с максимальной энтропией, то есть наиболее "неопределённой" в статистическом смысле функции. Величина δ соответствует величине погрешности в данных. Величина параметра стабилизации α определяется из условия $\|Y - U(\Psi)\|^2 = \delta$.

Проблема выбора величины параметра α является одной из центральных при решении некорректных задач и решается в зависимости от характера данных и имеющейся априорной информации о характере решения. Так, например, физические соображения и предварительные оценки могут указать интервал значений, которому принадлежит искомое решение и которому должно принадлежать построенное регуляризованное решение. Согласование величины параметра стабилизации с величиной погрешности данных при случайном характере погрешности составляет суть метода невязки [23]. При случайном характере наблюдений для выбора величины параметра α широко используется подход, основанный на выборе статистически наиболее надёжной модели [19, 24, 25].

Стабилизация оценки числа ВИЧ инфицированных

Рассмотрим алгоритм оценивания числа ВИЧ инфицированных исходя из интегрального уравнения (2) путём минимизации регуляризованного функционала (4) с выбором величины параметра регуляризации для построения статистически надёжной оценки [24]. Представим данные о зарегистрированном числе диагнозов СПИД и неизвестном числе ВИЧ инфицированных по когортам рождения: $U_i^k = U(t_i^k, x_i^k)$, $\Psi_j^k = \Psi(d^k + s_j, s_j)$, где k – номер рассматриваемой когорты, $d^k = t_i^k - x_i^k$ – параметр, зависящий лишь от номера когорты, $\{s_1, \dots, s_N\}$ – общая для всех когорт сетка разбиения по возрастам с шагом в 1 год. Для когорты k уравнение (2) в матричной форме принимает вид

$$U^k = A^k \Psi^k,$$

где $U^k = \{U_1^k, \dots, U_{n_k}^k\}$ и $\Psi^k = \{\Psi_1^k, \dots, \Psi_{m_k}^k\}$ векторы, а треугольная матрица A^k составлена из элементов

$$a_{ij}^k = \begin{cases} 0 & s_j > x_i^k \\ \beta(t_i^k, x_i^k) L(x_i^k, s_j) \exp\left(-\int_{s_j}^{x_i^k} \mu_c(d_k + \tau, \tau) d\tau\right) & s_j \leq x_i^k \end{cases}$$

Объединяя уравнения для разных когорт получим результирующее уравнение

$$U = A\Psi,$$

где $U = \{U^k\}$, $\Psi = \{\Psi^k\}$ - агрегированные векторы, матрица A - блочная диагональная матрица на диагонали которой расположены матрицы A_k .

В качестве стабилизирующего функционала можно принять какой либо из описанных ранее стабилизирующих функционалов. Для простоты рассмотрим здесь функционал по-когортной нормы приращений решения Ψ

$$\Omega(\Psi) = \sum_k \frac{1}{m_k} \sum_{j=2}^{m_k} (\Psi_j^k - \Psi_{j-1}^k)^2. \quad (5)$$

Межкогортные различия в этом случае не контролируются. Регуляризованный функционал невязки равен

$$\|Y - A\Psi\|^2 + \alpha\Omega(\Psi)$$

и принимает минимальное значение на решении

$$\Psi_\alpha = (A^T A + \alpha B)^{-1} A^T Y.$$

Матрица B - блочная диагональная состоящая из трёхдиагональных матриц вида

$$B^k = \begin{pmatrix} 1 & -1 & \Lambda & 0 & 0 \\ -1 & 2 & \Lambda & 0 & 0 \\ \Lambda & \Lambda & \Lambda & \Lambda & \Lambda \\ 0 & 0 & \Lambda & 2 & -1 \\ 0 & 0 & \Lambda & -1 & 1 \end{pmatrix}. \quad (6)$$

Вне блоков элементы матрицы нулевые. В [24] предложено выбирать величину параметра регуляризации из условия минимизации по α функционала

$$I_\alpha = \frac{\|Y - A\Psi_\alpha\|^2}{1 - \frac{2}{N} \text{tr}A(A^T A + \alpha B)^{-1} A^T}, \quad (7)$$

который оценивает среднее значение погрешности приближения невозмущённых данных при фиксированном α . Минимизация проводится по значениям α , при которых знаменатель в выражении для I_α положителен, N – число строк в матрице A . Выражение $\text{tr}C$ равно сумме диагональных элементов матрицы C . Критерий для выбора параметра регуляризации (7) сохраняется и при использовании других выражения для стабилизирующего функционала. В качестве матрицы B используется матрица из выражения для градиента стабилизирующего функционала.

Имитационное исследование алгоритма оценки числа ВИЧ инфицированных в популяции

Экспериментальное исследование эффективности описанного подхода для оценки числа лиц, инфицированных ВИЧ по заболеваемости СПИД, проводилось в два этапа. На первом этапе по заданным гипотетическим зависимостям вероятности инфицирования человека возраста x в году t $\psi(t, x)$ и по численности населения возраста x в году t $N(t, x)$ рас-

считывалось число людей $\Psi(t, x)$ возраста x инфицированных ВИЧ в году t как

$$\Psi(t, x) = \psi(t, x)N(t, x).$$

Затем, по заданному распределению длительности инкубационного периода $L(x, s)$ и смертности по всем причинам в возрасте x в году t $\mu_c(t, x)$ рассчитывалось число диагнозов СПИД в году t по формуле

$$U(t, x) = \int_0^x L(x, s) \exp\left(-\int_s^x \mu_c(t-x+\tau, \tau) d\tau\right) \Psi(t-x+s, s) ds \quad (8)$$

Полученная зависимость $U(t, x)$ выступает в качестве результатов регистрации числа диагнозов СПИД, смертность характеризует покогортное уменьшение популяции, а распределение длительности инкубационного периода описывает динамику развития заболевания, называемого СПИД.

При реализации вычислений рассматривалось дискретное представление непрерывных зависимостей в возрастном интервале 19-45 лет и 17 летнем временном интервале, соответствующем периоду 1986-2002, с шагом в один год. Применялись следующие формулы для вычисления числа диагнозов СПИД:

для когорт рождения 1941-1957 годов при $29+t-1986 < x \leq 45$, $1986 \leq t \leq 2002$

$$U(t, x) = \sum_{s=29+t-1986}^x L(x, s) \exp\left(-\sum_{\tau=s}^x \mu_c(t-x+\tau, \tau)\right) \Psi(t-x+s, s),$$

для когорт рождения 1957-1967 годов при $19+t-1986 < x \leq 29+t-1986$, $1986 \leq t \leq 2002$

$$U(t, x) = \sum_{s=19+t-1986}^x L(x, s) \exp\left(-\sum_{\tau=s}^x \mu_c(t-x+\tau, \tau)\right) \Psi(t-x+s, s),$$

для когорт рождения 1968-1983 годов при $19 < x \leq 19+t-1986$, $1986 \leq t \leq 2002$

$$U(t, x) = \sum_{s=19}^x L(x, s) \exp\left(-\sum_{\tau=s}^x \mu_c(t-x+\tau, \tau)\right) \Psi(t-x+s, s).$$

Все приведённые формулы легко представить в матричной записи

$$\bar{U} = A\bar{\Psi},$$

где \bar{U} и $\bar{\Psi}$ - значения числа диагнозов СПИД и числа инфицированных лиц за год для членов когорты рождения в году d . Элементы матрицы A вычисляются по формуле

$$A_{ij} = \begin{cases} 0 & s_j > x_i \\ L(x_i, s_j) \exp\left(-\sum_{\tau=s_j}^{x_i} \mu_c(d+\tau, \tau)\right) & s_j \leq x_i \end{cases}.$$

На втором этапе решалась обратная задача. То есть, по рассчитанной зависимости $U(t, x)$ и при заданных зависимостях $L(x, s)$ и $\mu_c(t, x)$ вычислялась оценка для числа ВИЧ инфицированных людей $\hat{\Psi}(t, x)$ (в формуле что-то лишнее? Где, в какой?), которая затем сравнивалась с исходной зависимостью $\Psi(t, x)$. Для членов одной когорты рождения нестабилизированная оценка числа ВИЧ инфицированных за год равна

$$\bar{\Psi} = (A^T A)^{-1} A^T \bar{U}.$$

В качестве стабилизирующего функционала при вычислениях принимался функционал покогортной нормы приращений решения (5), приводящий к выражению для стабилизированного решения в виде

$$\bar{\Psi}_\alpha = (A^T A + \alpha B)^{-1} A^T \bar{U},$$

матрица B – трёхдиагональная, вычисляется по формуле (6). Величина параметра регуляризации α вычислялась из условия минимизации критерия (7).

Результаты расчетов

Результаты оценки по описанной методике числа лиц, инфицированных ВИЧ, приведены на рисунках 1-5. Плотность распределения вероятности длительности инкубационного периода $L(x,s)$ задавалось в виде экспоненциальной зависимости со средним значением, эквивалентным 10 годам

$$L(x, s) = 0.1 \exp(-0.1(x - s)).$$

Смертность от всех причин в интервале возрастов 19-45 лет условно принималась равной 1% и не зависела от возраста и календарного года. Число случаев СПИД вычислялось по формуле (8) на основании гипотетической зависимости числа случаев инфицирования ВИЧ за год. В экспериментах использовалась одноэкстремальная зависимость

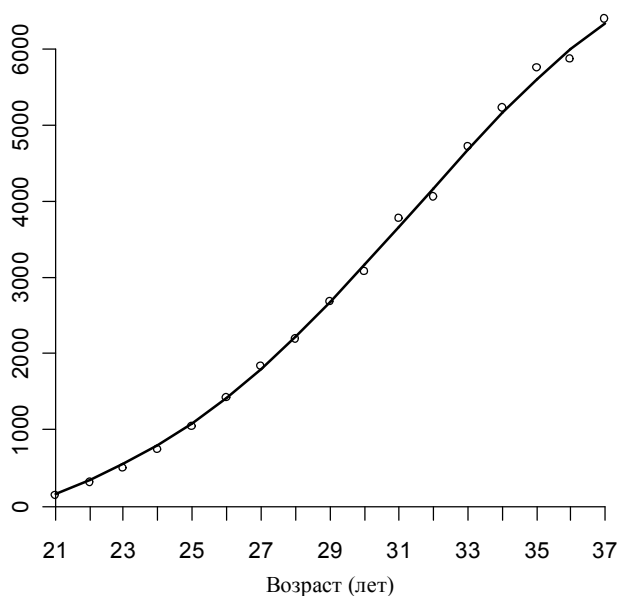
$$\Psi(t, x) = 10 \exp\left(-\left(\frac{x - 35}{10}\right)^2\right). \quad (9)$$

Относительная точность регистрации числа случаев СПИД принималась равной 5% и моделировалась с помощью генератора равномерно распределённых случайных чисел, которые искажали вычисленные значения числа случаев СПИД.

На рисунке 1 точками представлено ежегодное число случаев СПИД, смоделированное для когорты 1965 года рождения в возрастах с 1986 по 2002гг. Сплошная линия представляет результаты расчётов ежегодного числа случаев

СПИД по оценке числа ВИЧ инфицированных в этой когорте.

Рис 1. Число диагнозов СПИД, ежегодно зарегистрированных с 1986 по 2002гг. в когорте 1965 года рождения. Точки – симулированные данные, сплошная линия – результат оценки через оценку числа ВИЧ инфицированных.



На рисунке 2 представлены оценки числа ежегодно инфицированных ВИЧ людей из той же когорты. Точки соответствуют числу ежегодно инфицированных ВИЧ людей, рассчитанных по числу диагнозов СПИД без регуляризации. Видно, что даже при малых искажениях кривой числа диагнозов СПИД, представленной на рисунке 1, оценка числа ВИЧ инфицированных за год демонстрирует высокую степень нестабильности. Для стабилизации решения использовался метод регуляризации с оптимальным выбором вели-

чины параметра регуляризации по критерию (7). Соответствующее решение представлено на рисунке 2 сплошной линией. Для сравнения на рисунке пунктиром показана точное число ежегодно инфицированных ВИЧ людей, использованное при симуляции числа случаев СПИД на рисунке 1. Видно, что предложенный метод позволил стабилизировать оценку без существенной потери точности.

Рис. 2. Число лиц, ежегодно инфицированных ВИЧ в период с 1986 по 2002гг. в когорте 1965 года рождения. Точки – оценка по числу диагнозов СПИД без регуляризации, сплошная линия – оценка по числу диагнозов СПИД с оптимальным выбором величины параметра регуляризации, пунктир – теоретическое число ежегодно инфицированных ВИЧ, использованное для симуляции числа случаев СПИД.

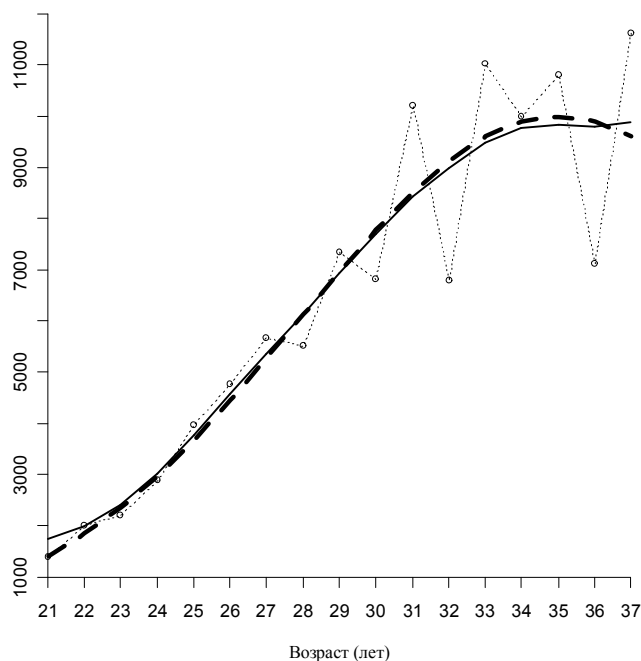
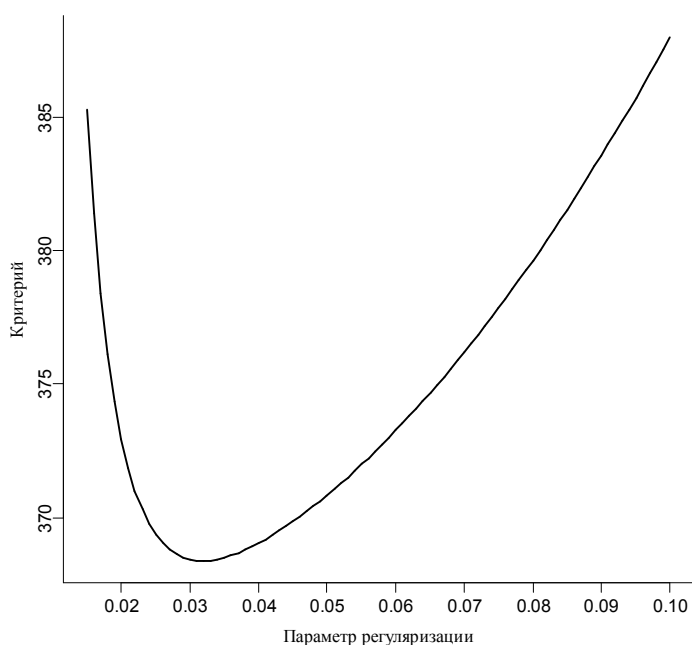


Рис. 3 иллюстрирует процедуру выбора оптимальной величины параметра регуляризации, которая соответствует минимальному значению критерия.

Рис. 3. Зависимость величины критерия (условные единицы) от величины параметра регуляризации.



На рис. 4 представлены оценки числа ежегодно инфицированных ВИЧ людей из когорты 1955 года рождения. Точки соответствуют числу ежегодно инфицированных ВИЧ людей, рассчитанных по числу диагнозов СПИД без регуляризации. Сплошной линией показана оценка с применением оптимальной величины параметра регуляризации, пунктиром - исходное число ВИЧ инфицированных за год, вычисленное по модели (9). На рисунке 5 представлены аналогичные зависимости для когорты 1975 года рождения.

Рисунок 4. Число лиц, ежегодно инфицированных ВИЧ в период с 1986 по 2002гг. в когорте 1955 года рождения. Точки – оценка по числу диагнозов СПИД без регуляризации, сплошная линия – оценка по числу диагнозов СПИД с оптимальным выбором величины параметра регуляризации, пунктир – теоретическое число ежегодно инфицированных ВИЧ, использованное для симуляции числа случаев СПИД.

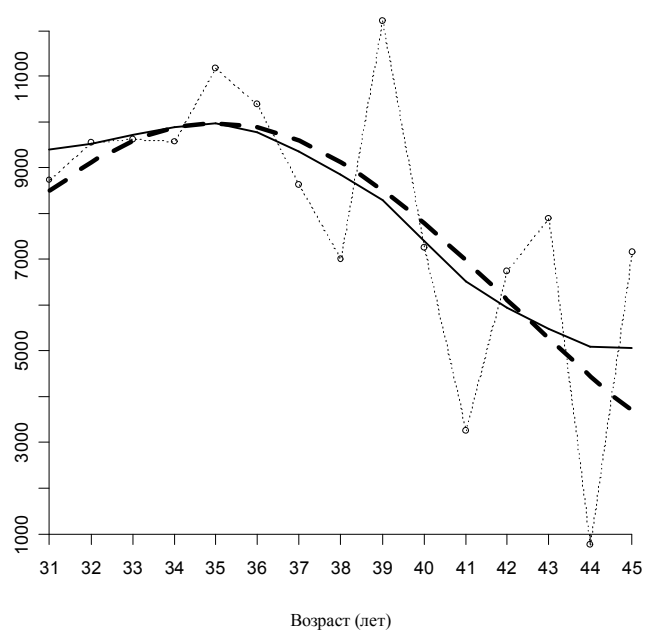
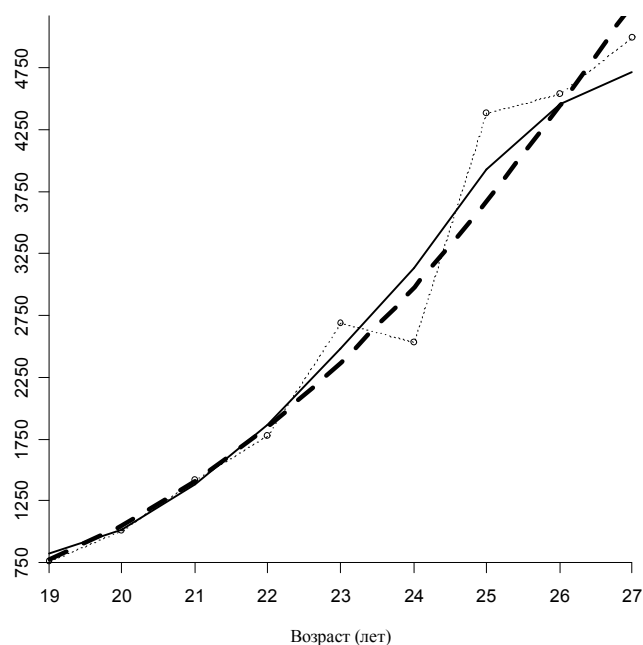


Рис. 5. Число лиц, ежегодно инфицированных ВИЧ в период с 1986 по 2002гг. в когорте 1975 года рождения. Точки – оценка по числу диагнозов СПИД без регуляризации, сплошная линия – оценка по числу диагнозов СПИД с оптимальным выбором величины параметра регуляризации, пунктир – теоретическое число ежегодно инфицированных ВИЧ, использованное для симуляции числа случаев СПИД.



Заключение

Уточнение числа ВИЧ инфицированных в группах риска важно как для оптимизации необходимой текущей медицинской помощи, так и для построения надёжного прогноза ожидаемого развития эпидемии ВИЧ/СПИД. По последним опубликованным данным в США около 250 тысяч человек, инфицированных ВИЧ, не учтены в общей статистике, не получают лечение и являются неконтролируемым источником распространения вируса [26]. В Российской Федерации существует развитая система контроля эпидемии ВИЧ/СПИД, что позволяет надеяться на более высокий процент выявляемости случаев заражения, однако надёжность получаемой информации требует независимой оценки, для чего могут быть использованы математические ме-

тоды оценки инфицированности ВИЧ по динамике роста числа диагнозов СПИД.

Проблема оценки числа ВИЧ инфицированных по числу больных СПИД является сложной задачей обработки данных. Основная трудность заключается в наличии длительного инкубационного периода между моментом заражения и началом развития заболевания. Это приводит к тому, что диагноз СПИД может быть обусловлен фактом заражения ВИЧ на длительном временном периоде. С другой стороны, наличие длительного инкубационного периода приводит к сильным искажениям в оценках числа ВИЧ инфицированных за год, что требует привлечения специальных методов оценки. Рассмотренный в статье метод стабилизации оценки числа инфицированных ВИЧ с автоматическим выбором величины параметра регуляризации позволяет повысить достоверность оценки, что продемонстрировано в модельных расчётах. Для более точной оценки числа ВИЧ инфицированных необходимо уточнить закон распределения длительности инкубационного периода, который, теоретически, зависит от множества факторов и, в частности, от пути проникновения в организм человека. Для такого уточнения необходимо использовать клинические и эпидемиологические данные развития СПИД в разных группах риска.

Литература

1. Piot P., Bartos M., Ghys P.D., Walker N., Schwartlander B. The global impact of HIV/AIDS. *Nature* 2001, 410: 968-973
2. Денисов Б.П. Определение численности групп населения, уязвимых для ВИЧ. Тезисы международной конференции "ПОЛИТИКА НАРОДОНАСЕЛЕНИЯ: НАСТОЯЩЕЕ И БУДУЩЕЕ - ЧЕТВЕРТЫЕ ВАЛЕНТЕЕВСКИЕ ЧТЕНИЯ", Москва 2005, с.164

3. Tedstrom J., Narkevich M. Why is HIV/AIDS a Public Policy Priority for Russia? Transatlantic Partners Against AIDS, 2004.
4. Грисин С.А., Уоллендер С.А. ВИЧ.СПИД Кризис в России: Противостояние настоящему и лицом к лицу с будущим. CSIS HIV/AIDS TASK FORCE, 2002.
5. Mariotto B., Verdecchia A. Using AIDS mortality data to reconstruct HIV/AIDS epidemics. *Statistics in Medicine* 2000, 19: 161-174
6. Rino Bellocco, Ian C. Marschner Joint analysis of HIV and AIDS surveillance data in back-calculation *Statistics in Medicine* 2000, 19: 297-311
7. Artzrouni M. Back-calculation and projection of the HIV/AIDS epidemic among homosexual/ bisexual men in three European countries: evaluation of past projections and updates allowing for treatment effects. *Eur J Epidemiol* 2004, 19:171-179
8. Gigli A., Verdecchia A. Uncertainty of AIDS incubation time and its effects on back-calculation estimates. *Statistics in Medicine* 2000, 19: 175-189
9. Munoz A., Xu J. Models for the incubation of AIDS and variations according to age and period. *Statistics in Medicine* 1996; 15: 2459-2473
10. Solomon P.J., Wilson S.R. Accommodating change due to treatment in the method of back projection for estimating HIV infection incidence. *Biometrics* 1990, 46, 1165-1170
11. Becker N.G., Motika M. Smoothed nonparametric back-projection of AIDS incidence data with adjustment for therapy. *Mathematical Biosciences* 1993; 118: 1-23
12. Deuffic S., Costagliola D. Is the AIDS incubation time changing? A back-calculation approach. *Statistics in Medicine* 1999, 18: 1031-1047
13. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. М.: Наука, 1974.
14. Bacchetti P., Segal M.R., Jewell N.P. Backcalculation of HIV Infection Rates *Stat. Sci.* 1993, 8: 82-101

15. Brookmeyer R., Gail M.H. A Method for Obtaining Short-Term Projections and Lower Bounds on the Size of the AIDS Epidemic. *JASA* 1988, 83: 301-308
16. Bellocco R., Pagano M. Multinomial analysis of smoothed HIV back-calculation models incorporating uncertainty in the AIDS incidence. *Statistics in Medicine* 2001, 20: 2017-2033
17. Cazelles B. and Chau N.P. Using the Kalman Filter and Dynamic Models to Assess the Changing HIV/AIDS Epidemic. *Mathematical Biosciences* 1997, 140: 131-154
18. Liao J, Brookmeyer R. An empirical Bayes approach to smoothing in backcalculation of HIV infection rates. *Biometrics* 1995, 51: 579-588
19. Вапник В.Н., Глазкова Т.Г., Кощев В.А., Михальский А.И., Червоненкис А.Я. Алгоритмы и программы восстановления зависимостей. М.: Наука, 1984
20. Stone M. Cross-Validation Choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Ser. B* 1974, 36, 111-147
21. Tan W. Stochastic modelling of AIDS epidemiology and HIV pathology. World Scientific Publishing Co., 2000
22. Amato U., Hughes W. Maximum entropy regularization of Fredholm integral equations of the first kind. *Inverse Problems* 1991, 7: 793-808
23. Морозов В. А. Регулярные методы решения некорректных задач. - М.: Изд-во МГУ, 1974.
24. Михальский А.И. Выбор алгоритма оценивания по выборкам ограниченного объема. *АиТ* 1987, 7, 91-102.
25. Michalski A.I., Yashin A I (1989) Analysis of AIDS epidemic as an inverse problem of population studies. Paper presented at the Workshop on Modelling the spread of HIV/AIDS and its Demographic and Social Consequences. Budapest, Hungary, 23-24 Nov, organized by IIASA, 1989.
26. CDC. Use of Social Networks to Identify Persons with Undiagnosed HIV Infection - Seven U.S. Cities, October 2003-September 2004. *MMWR* 2005;54:601-605.