СТАТИСТИЧЕСКАЯ ОЦЕНКА ВЕРОЯТНОСТИ ЗАРАЖЕНИЯ ВИЧ-ИНФЕКЦИЕЙ НА ОСНОВЕ ДАННЫХ О ПОСЛЕДНИХ ЭПИЗОДАХ РИСКОВАННОГО ПОВЕДЕНИЯ

А. Е. Пащенко[1], А. Л. Тулупьев[2], С. И. Николенко[3]

[1,2]Санкт-Петербургский институт информатики и автоматизации РАН, [3]Санкт-Петербургское отделение Математического института им. В. А. Стеклова РАН

[1,2]СПИИРАН, 14-я линия ВО, д. 39, Санкт-Петербург, 199178, [3]ПОМИ РАН, Санкт-Петербург, наб. р. Фонтанки, д. 27, 191023

[1]<pae_82@mail.ru>, [2]<alt@iias.spb.su>, [3]<sergey@logic.pdmi.ras.ru>

УДК 681.3

Пащенко А. Е., Тулупьев А. Л., Николенко С. И. **Статистическая оценка вероятности заражения ВИЧ-инфекцией на основе данных о последних эпизодах рискованного поведения** // Труды СПИИРАН. Вып. 3, т. 2. — СПб.: Наука, 2006.

Аннотация. *Предлагается постановка задачи по оценке вероятности заражения ВИЧ-инфекцией, опирающейся на сведения о нескольких последних эпизодах рискованного поведения. В работе описана математическая модель заражения инфекционными заболеваниями, позволяющая корректно поставить математическую задачу обработки данных о последних эпизодах. Предложены пути решения этой задачи.* — Библ. 12 назв.

UDC 681.3

Paschenko A. E., Tulupyev A. L., Nikolenko S. I. **HIV-Acquisition Risk Statistical Estimates Based on the Data about Several Last Episodes of Risky Behavior** // SPIIRAS Proceedings. Issue 3, vol. 2. — SPb.: Nauka, 2006.

Abstract. *We offer a task statement on estimating HIV-acquisition risk. Our method is based on using data concerning several last episodes of risky behavior. We describe a mathematical model of acquiring infectious diseases that allows to correctly set up a mathematical task of considering last episodes data. We offer several approaches to solving this task.* — Bibl. 12 items.

1. Введение

В последние годы число заболевших вирусом ВИЧ в нашей стране продолжает неуклонно расти [1]. В то же время интенсифицировались исследования в области вирусологии в целом, мониторинга и профилактики данного заболевания [10]. Появились специальные программы по изучению возможностей профилактики, ряд исследований был проведен Биомедицинским центром Санкт-Петербурга [2]. В силу особенностей ВИЧ-инфекции (неизлечимость, высокая социальная и социально-эпидемиологическая опасность) она стала одной из наиболее серьезных проблем современного здравоохранения не только в Российской Федерации, но и в мире в целом. В некоторых странах Африки (например, в ЮАР), есть города, в которых не осталось никакого вида бизнеса, кроме похоронного [8]. Там престарелые женщины и мужчины вынуждены содержать и воспитывать своих внуков, поскольку среди людей среднего возраста СПИДом поражено более половины [7]. В России сейчас ситуация не настолько серьезная, но, по мнению авторов доклада [9], она сопоставима с ситуацией в странах Африки десятилетней давности и может привести к таким же последствиям, если не предпринять своевременных шагов.

Существующие масштабы эпидемии ВИЧ и связанные с ней угрозы, ставят ряд новых междисциплинарных задач. Среди них:

1) измерение вероятности (более обще — оценка риска) заразиться ВИЧ-инфекцией для индивида в популяции или локальной группе;

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

2) сравнение рисков между двумя популяциями (в одно время);
3) сравнение рисков в одной и той же популяции («до и после» некоторого события или процесса);
4) оценка эффективности превентивных программ по изменению риска заразиться (программы снижения вреда);
5) оценка эффективности превентивных программ по стоимости;
6) мониторинг и своевременное обнаружение «скачков» риска.

Таким образом, для решения стоящих перед российским здравоохранением неотложных проблем требуется искать методы решения вышеприведенных задач. В настоящей работе мы составим обзор современных подходов к измерению риска заразиться ВИЧ-инфекцией, основанных на самоотчетах респондентов, а также связанных с этим теоретических и практических затруднений. К сожалению, выполнение прямых измерений риска представляется нереалистичным (слишком дорого и труднореализуемо на практике). Поэтому необходимо использовать косвенные измерения. Это значит, что потребуется предложить математическую модель для представления и обработки данных косвенных измерений, которая бы адекватно отражала полученные нами сведения и делала на их основании разумные выводы. Настоящая работа посвящена постановке исследовательской задачи на разработку одной из таких моделей — модели, основанной на учете последних эпизодов рискованного поведения.

2. Социально-эпидемиологическая терминология

Чтобы лучше понять социальные и эпидемиологические составляющие данной проблемы, необходимо изложить систему определений и показателей, использующихся для характеристики эпидемии ВИЧ-инфекции.

Прежде всего необходимо разграничить три понятия, связанных с этой инфекцией: ВИЧ, ВИЧ-инфекция и СПИД.

ВИЧ — это вирус иммунодефицита человека, который является возбудителем болезни. Наиболее распространенный в употреблении термин — *ВИЧ-инфекция* — на самом деле подразумевает болезнь, вызванную вирусом. Эта болезнь имеет несколько стадий, которые классифицированы и рассмотрены в специальной литературе. Последняя стадия данной болезни и называется СПИДом.

СПИД — синдром приобретенного иммунодефицита. Разберем эту аббревиатуру более подробно:

1) *синдром* — совокупность признаков и симптомов данного заболевания;
2) *приобретенного* — генетически не обусловленного, а полученного в процессе жизни индивида;
3) дефицит — недостаток, в данном случае в работе иммунной системы, *иммунодефицит* — поражение иммунной системы, ее неспособность противостоять инфекциям.

Также необходимо более подробно рассмотреть понятие иммунной системы, поскольку именно она подвергается воздействию ВИЧ**.** Иммунитет — это особая функция организма человека защищаться от живых тел и веществ, несущих на себе признаки генетически чужеродной информации. Иммунная система вырабатывает специфические молекулы — антитела — для борьбы с различными возбудителями и чужеродными веществами (антигенами). Для ВИЧ-инфекции одними из таких тел являются тела «CD4». При проникновении в организм чужеродных агентов (вирусов, бактерий) включается иммунный от-

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

вет, в котором участвуют особые клетки крови — лимфоциты**.** Лимфоциты распознают возбудителей, блокируют их разрушительное действие и уничтожают их, а также способствуют выработке антител [11].

В исследованиях говорят о *риске заразиться ВИЧ-инфекцией*. Разделяют два похода к изучению этого понятия:

а) HIV Infection Acquisition Risk (риск приобретения ВИЧ-инфекции); в этом подходе рассматривается только потенциальный реципиент; объект исследования — один представитель рисковой группы, например инъекционный наркопотребитель;

б) HIV Infection Transmission Risk (риск передачи ВИЧ-инфекции); этот подход предоставляет заметно более глубокий анализ процесса распространения ВИЧ; в нем рассматривается как потенциальный реципиент, так и носитель инфекции, который может ею заразить; пример объекта исследования — дискордантная пара, в которой риски заразиться между разными партнерами могут быть несимметричны.

Также необходимо определить, что же мы понимаем под термином *риск* применительно к данной проблеме в математическом смысле. В общем случае под *риском заражения* понимается вероятность заражения. Но, учитывая неизбежное расширение понятия, под этим же названием также используется и соотношение вероятностей (odds ratio) — отношение инсиденс-показателей (см. ниже). А учитывая сложность получения вероятности в «чистом виде», под риском заражения иногда понимают любые сведения, количественные и даже качественные, характеризующие эту вероятность. Требуется научиться обрабатывать такие данные и делать на их основании разумные, математически обоснованные выводы.

3. Способы измерения риска заразиться

Как правило, один и тот же процесс можно описывать несколькими параметрами, причем разными способами. Основными величинами, характеризующими масштабы эпидемии ВИЧ, являются преваленс-показатель (*prevalence*) и инсиденс-показатель (*incidence*) [12].

Преваленс-показатель [12] — это показатель пораженности инфекцией. Математически это отношение больных людей к общему числу людей в популяции в заданный момент времени. Данный показатель формально является безразмерным; фактически же рассматриваются случаи заболевания, например, на 100 тысяч человек, на 10 тысяч человек и т.д.

Преваленс-показатель — один из наиболее используемых показателей для различного рода измерений. Поэтому вполне естественно, что существуют различные его «подвиды». Он широко используется для характеристики риска заразиться, эффективности работы системы здравоохранения.

Инсиденс-показатель [12] — это частота заболеваемости за определенный период времени. Математически это отношение заболевших людей из заданной группы, *находящейся под риском заболеть*, в заданный период, к общему количеству человеко·лет, «накопленных» за период наблюдения в этой группе. Его размерность — $\frac{\text{случаи}}{\text{человеко}\times\text{годы}}$ (cases per person per year).

Для инсиденс-показателя существуют достаточно строгие и глубоко проработанные процедуры организации измерений и их последующих расчетов [12].

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Следует отметить, что использование преваленс-показателя совершенно неприменимо к проблеме оценки риска заразиться ВИЧ-инфекцией. Данная болезнь имеет ряд особенностей, и одна из основных — неизлечимость больных. Яркой иллюстрацией может послужить следующий пример. Предположим, что есть две популяции: в первой преваленс-показатель — 20%, инсиденс-показатель — 0%, а во второй оба показателя составляют по 10%. В первом случае можно говорить о полном «замораживании» распространения инфекции. Во втором показатель инсиденс настолько велик, что уже совсем скоро ВИЧ-инфекцией будут поражены почти все люди, находящиеся под риском заразиться в данной популяции.

Инсиденс-показатель является золотым стандартом характеристики риска заразиться. Казалось бы, он должен широко использоваться, но существует ряд серьезных проблем со сложностью его измерения и дороговизной полученных знаний.

Прямые измерения инсиденс-показателя представляют собой когортное исследование. Данный тип исследования имеет следующие общие черты: случайным образом выбирается некоторое количество людей (допустим, оно составляет 500–1000 человек, — их количество зависит от предполагаемых результатов исследования). Сразу следует отметить, что если такую выборку производить из общей популяции людей, то число вновь заразившихся ВИЧ-инфекцией за определенный период в среднем составит всего несколько человек, и в данной выборке может даже не найтись ни одного вновь заразившегося. Следовательно, по результатам исследования будет невозможно сделать какие-либо выводы. Поэтому данное количество людей выбирается из определенной группы очень высокого риска, например людей употребляющих наркотики либо других групп (IDU, MSM, HET, SW и т.д. — см. ниже).

Далее выбираются люди, удовлетворяющие требованиям исследования. Критериями выбора могут являться, например, частота и способ употребления наркотиков, а также уровень «рискованности» поведения данного человека. Затем всех этих людей сопровождают[1] в течение некоторого периода времени, обычно от года до полутора. Это сопровождение, как и просто удержание человека в исследовании — очень сложная задача. Как правило, исследование в целом — подготовка, активная фаза исследования, которая обычно включает в себя несколько этапов и несколько групп участников, обработка и оценка полученных результатов, — занимают до 5–7 лет. Также требуется соблюсти этические нормы. И, наконец, финансирование, требуемое для данного вида исследования, исчисляется обычно миллионами долларов, причем проведение подобного исследования не всегда возможно даже в принципе.

Сокращать затраты совершенно необходимо: данные еще и быстро устаревают. Даже небольшая относительная экономия при измерениях будет очень заметна в абсолютных величинах, т.к. финансовые затраты весьма масштабны. Подобная ситуация для статистики не нова. В таких случаях, когда нельзя или сложно измерить напрямую, обычно прибегают к *косвенным измерениям*. Как правило, за косвенными измерениями стоит определенная математическая модель интересующего нас объекта или процесса.

[1]Под сопровождением понимают не буквальное «сопутствование» участнику исследования, куда бы он ни пошел, а комплекс мер, направленный на своевременное прохождение этим участником процедур исследования; например, звонок по телефону за несколько дней до даты следующей запланированной процедуры.

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

4. Математическая модель косвенных измерений

Основополагающей для данной модели можно считать статью «Modeling HIV Risk» [5] — «Моделирование риска заражения ВИЧ». Основные параметры модели, предложенной в этой статье таковы:

1) имеется некоторое число K видов рискованного поведения. При этом множество видов рискованного поведения охватывает все мыслимые способы передачи ВИЧ-инфекции: поцелуи, различные виды половых контактов, внутривенный общий прием наркотиков, использование общей посуды для их приготовления и другие виды рискованного поведения. Следует учитывать, что для различных участников полового контакта вероятность заразиться за один эпизод рискованного поведения будет различной, порой очень существенно;
2) известны вероятности p_k заразиться за отдельный эпизод конкретного вида рискованного поведения при заданном способе участия. С целью установить эти вероятности был проведен целый ряд биологических исследований, которые достигли успеха: их результатами стали вероятности передачи инфекции за один эпизод рискованного поведения;
3) предполагается, что можно подсчитать число эпизодов рискованного поведения у данного человека за интересующий нас период времени для каждого его вида рискованного поведения. Т.е. получить число (N_i) эпизодов каждого вида рискованного поведения при заданном способе участия. Оценка этих чисел и является основным предметом последующего исследования — остальные компоненты уже имеются априори;
4) предполагаются также некоторые независимости; в частности, возможности заражения разными путями предполагаются независимыми.

При известной вероятности p_i передачи ВИЧ-инфекции за один эпизод рискованного поведения i-того вида, а так же при известном числе эпизодов N_i в которых участвовал респондент, вероятность передачи ВИЧ инфекции составит $\mathrm{Pr}_i = 1-(1-p_i)^{N_i}$; если выявить все виды рискованного поведения, то общая оценка риска будет $\mathrm{Pr} = 1-\prod_{i=1}^{n}(1-\mathrm{Pr}_i)$.

Рискованное поведение классифицируют по способам передачи:

1) сексуальные контакты;
2) общее употребление наркотиков:
    а) общая посуда,
    б) общий шприц,
    в) общие фильтры (вата).

Существуют три основные группы с повышенным уровнем риска передачи инфекции:

1) IDU — Injecting Drug User (инъекционные наркопотребители);
2) MSM — Men who had Sex with Men (мужчины, у которых был секс с мужчиной);
3) HET — HETerosexual partners (гетеросексуальные партнеры).

У каждой из этих групп преобладают свои виды рискованного поведения, детальная классификация которых уже составлена.

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

5. Оценка числа эпизодов

Для практической реализации представленной математической модели требуется измерить количество эпизодов рискованного поведения за интересующий исследователя период времени. Для этого нужно связать наблюдаемые параметры (ответы респондентов в опросниках риска) с величинами, которые мы действительно хотим измерить.

Сейчас для этого используются два метода, каждый из которых имеет недостатки. Первый метод — прямые (даже, скорее, прямолинейные) вопросы: «Сколько раз Вы *делали так* в течение последнего месяца (трех, шести, года)?». На такие вопросы респонденты обычно дают практически не соотносящиеся с реальностью ответы. Действительно, можно задать себе вопрос: «Сколько раз за последние три месяца я пил чай с сахаром?» или «Сколько бананов я съел за последние полгода?». Попытка ответа даже самому себе даст четкую картину незначительной достоверности такого ответа.

Второй метод — Лайкерт-шкалы [4] — опросники, в которых используются качественные, а не количественные варианты: «Никогда», «Редко», «Иногда», «Часто», «Всегда» и подобные им возможности для ответа. Вопрос ставится легко, ответ тоже получить несложно, однако эти ответы не несут никаких полезных сведений относительно числа эпизодов: то, что «Часто» для одного человека, может быть «Редко» для другого, а то, что «Часто» в одном виде поведения, может быть «Редко» для другого вида поведения. Кроме того, «расстояние» между «Всегда» и «Очень часто» совершенно не обязательно совпадает с расстоянием между «Редко» и «Никогда». На практике шкалы арифметизируют, но за этой арифметизацией не стоит никакой достоверной гипотезы; получающиеся расчеты ситуацию с риском не характеризуют вообще никак. Таким образом, возникает потребность в более адекватных источниках сведений о рискованном поведении и методиках их обработки, которые сделают возможной более обоснованную оценку числа эпизодов.

Одной из возможных альтернатив Лайкерт-шкал представляется опрос респондента об одном или нескольких последних эпизодах рискованного поведения. Такой опрос позволяет судить об интервалах между эпизодами рискованного поведения, а так же об интервале между временем опроса и последним эпизодом.

Уже сами упомянутые интервалы могут оказаться более удобным косвенными оценками риска, чем порядковые шкалы вида «Никогда, Редко, Иногда, Часто, Всегда». Чем меньше интервалы, тем более высокую степень риска мы можем ожидать. Интервалы можно сравнивать между собой. Зачастую респонденты, особенно люди со сложным социальным положением, хотят, чтобы у интервьюера сложилось о них положительное впечатление, и сознательно дают заведомо ложные ответы; иногда они могут запутаться или ошибиться при ответе. Данная проблема, на самом деле, является одной из наиболее острых при проведении исследования. Предлагаемый нами подход выводит на новый уровень ее решение: во время интервью программное приложение может сравнивать зависимые либо вложенные ответы на вопросы и выдавать об этом сообщение интервьюеру.

Существуют еще ряд преимуществ данного подхода:

1) ответы не надо арифметизировать, они и так количественные, более того — континуальные;

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

2) чем короче интервалы, тем более высокую степень риска, как уже было сказано, мы можем предполагать;
3) интервалы можно сравнивать между собой;
4) можно рассчитывать коэффициенты корреляции — данный инструмент наиболее востребован в гуманитарных науках для поиска зависимостей как между ответами на отдельные вопросы, так и на группы вопросов.

Следовательно, задачей опроса должно стать определение величины N_i, т.е. социологический опрос должен был бы обеспечить сбор таких данных о серии последних эпизодов, которые позволили бы в дальнейшем рассчитать корректную количественную оценку риска. Но на практике существует ряд проблем при реализации данного подхода.

Во-первых, существует нечеткость в ответах. Формулировки ответов в большинстве случаев строятся из следующих слов [2]:

— сегодня [утром | днем | вечером];
— [поза]вчера [утром | вечером];
— тот же день, [еще] неделю назад, [еще] две недели назад;
— в пятницу, в предыдущую пятницу.

Например:

— «сегодня, сегодня, вчера»;
— «сегодня утром, вчера вечером, вчера утром».

Также существует проблема недоопределенности ответов, которая выражает естественную неточность ответов. Пусть даны ответы двух респондентов об их последнем употреблении, например, чая. Один из них ответил, что он пил чай последний раз полгода назад, а другой назвал точную дату, например 06.03.05 (пусть опрос проходил 06.09.05). Совершенно понятно, что в первом случае «погрешность» или «ошибку» ответа правдоподобно было бы оценить как месяц, а во втором случае речь идет о «погрешности» величиной в сутки.

Следует более подробно рассмотреть, какое же на самом деле мы получаем значение. Фактически мы получаем не *точное* значение оценки времени между эпизодами интересующего нас вида рискованного поведения, а *интервальное* значение оценки времени, прошедшего между эпизодами. А если речь идет не об одном, а о нескольких ответах, то интервальная неопределенность оценки эпизодов одного респондента может оказаться больше, чем вся оценка времени между эпизодами другого респондента.

Существует ряд математических подходов для решения данной задачи:

1) можно брать среднюю точку интервала и проводить все дальнейшие расчеты с учетом того, как все значения положены на единую временную шкалу;
2) работать с интервальной оценкой;
3) использовать нечеткость (по Заде);
4) предположить некоторое вероятностное распределение.

Возможны и другие подходы; в дальнейшем планируется рассмотреть их применимость и сложность.

Еще одна существенная проблема — о каком количестве эпизодов интересующего исследователя поведения спрашивать. Этот вопрос является сложным из-за огромного количества факторов, оказывающих влияние на максимальную полноту и правдоподобность данных. Но основные «за» и «против» числа эпизодов изложены ниже.

Теоретически оптимально производить измерения и получать ответы для наибольшего N — числа эпизодов рискованного поведения, однако на практике

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

существует ряд ограничений при получении и обработке большого количества ответов. В случае, когда мы получаем ответы о трех последних, эпизодах можно строить различные распределения, в отличие от случая, когда нам известны ответы только о двух эпизодах. Скорее всего, при ответах о трех последних эпизодах получается самая высокая «правдоподобность» оценки: меньшее число эпизодов несет меньше сведений, а о большем числе эпизодов респондент начнет давать преднамеренно неверные ответы, ввиду сильной усталости и желания закончить опрос быстрее.

Исследования в данном направлении велись, в частности, Биомедицинским центром в рамках проекта изучения полового пути распространения ВИЧ-инфекции. В опросник одного из пилотных проектов был внедрен соответствующий блок вопросов [2]. Респонденту задавались вопросы о датах трех последних случаев внутривенного употребления наркотиков, трех последних сексуальных контактов и трех последних сексуальных контактов с использованием презерватива.

Результаты исследования [2] подтвердили работоспособность и применимость данного метода. Всего было опрошено 50 наркопотребителей. Только несколько из них не подходили по критериям опроса: не употребляли наркотики внутривенно и в интересующий исследователей период. Существенным является то, что только один наркопотребитель не смог дать ответ о третьем от момента опроса эпизоде: его ответ был «май 2004, минус пол года, не помню». Остальные 44 человека дали ответы о трех последних эпизодах внутривенного употребления наркотиков [2].

Конечно, чем больше помнит респондент, тем лучше. Возможно, следует рассмотреть ситуацию, когда сведения собираются до тех пор, пока респондент уверенно отвечает. Не следует забывать, что инъекционные наркопотребители — особый слой людей, и для них может представлять серьезную трудность вспомнить достаточно большое количество эпизодов интересуемого исследователей поведения. Математический аппарат, разумеется, рационально будет развивать для общего случая «небольшого N».

Данные о последних эпизодах можно пополнить другой информацией, которую респонденты также часто готовы предоставить:

1) максимальное и минимальное расстояние между эпизодами рискованного поведения;
2) обычный, с точки зрения респондента, интервал между эпизодами рискованного поведения.

На самом деле данные можно и нужно пополнять любыми доступными дополнительными сведениями, которые способны помочь при построении математической модели. Но не следует забывать, что опрос не должен быть чересчур длительным и тяжелым, иначе мы рискуем потерять внимание респондента, что может привести не только к отсутствию информации, но и (что еще хуже) к заведомо ложным ответам.

6. Математическая модель: индивидуальный подход

В этом разделе будет рассмотрен ряд возможных статистических и теоретико-вероятностных подходов к вычислению риска посредством аппарата последних эпизодов. Прежде всего следует отметить, что задача сразу естественным образом делится на две:

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

1) дана некоторая информация о нескольких последних эпизодах рискованного поведения *одного* отдельно взятого человека; требуется оценить риск заражения для данного индивидуума;
2) дана информация о последних эпизодах рискованного поведения некоторой статистически значимой выборки; требуется оценить *средний* риск заражения в данной популяции.

В данном разделе мы более подробно остановимся на обсуждении первой из этих задач, а второй посвятим следующий раздел.

Как моделировать эпизоды рискованного поведения? В первую очередь хочется рассмотреть рискованное поведение как *случайный процесс*. Задача ставится следующим образом: найти модель поведения данного испытуемого и оценить его риск заражения за интересующий нас период времени, которая бы являлась наиболее подходящей для решения данной задачи. Здесь мы предложим две возможные модели, однако, разумеется, могут потребоваться и другие варианты.

Наверное, самый классический и часто используемый в статистике случайный процесс — это пуассоновский процесс. Он возникает, когда вероятности попадания событий в непересекающиеся временные интервалы независимы. Пуассоновский процесс полностью характеризуется одним параметром, обычно обозначаемым через λ; вероятность того, что в интервале $[t_0, t_0 + t]$ произойдут k событий, равна

$$\Pr[N([t_0, t_0 + t]) = k] = \frac{e^{-\lambda t}(\lambda t)^k}{k!}.$$

Отметим, что для Пуассоновского процесса $\mathrm{E}\,X = \mathrm{D}\,X = \mathrm{E}(X - \lambda)^3 = \lambda$.

Пуассоновский процесс очень часто применяется на практике для моделирования различных случайных процессов. Однако иногда оказывается, что условие равенства первых трех моментов слишком жестко регламентирует пуассоновскую модель: возникает ситуация, когда дисперсия больше предписанного значения (overdispersion). Чтобы бороться с этой ситуацией, обычно считают, что пуассоновский параметр λ не фиксирован, а сам является случайной величиной. Чаще всего считают, что λ имеет гамма-распределение с параметрами μ, τ: плотность распределения λ имеет вид

$$\Gamma(\tau)^{-1}\left(\frac{\tau}{\mu}\right)^{\tau}\lambda^{\tau-1}\exp\left(\frac{-\lambda\tau}{\mu}\right),$$

а плотность исходного (уже не пуассоновского) распределения имеет вид

$$p(X = x) = \frac{\Gamma(x + \tau)}{x!\,\Gamma(\tau)}\left(\frac{\tau}{\mu + \tau}\right)^{\tau}\left(\frac{\mu}{\mu + \tau}\right)^{x}.$$

Существуют и другие подходы к моделированию случайных процессов. Задача статистических исследований на данном этапе состоит в том, чтобы выработать алгоритмы выбора того или иного распределения для моделирования, а затем подсчета его максимально правдоподобных параметров. Задача осложняется тем, что в индивидуальном случае данных *очень* мало; по статистическим меркам — почти нет совсем. Поэтому доверительные интервалы рассчитанных параметров, а, как следствие, и риска, в любом случае будут очень большими. Чтобы сделать более точные выводы, потребуется привле-

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

кать дополнительные данные, дополнительные гипотезы. Как будут заданы эти данные, как их учитывать, можно ли будет на их основании делать разумные оценки — покажут дальнейшие исследования.

7. Математическая модель: среднее по популяции

Предположим, что в исследовании участвует некоторая достаточно большая выборка из интересующей нас популяции, и мы хотим оценить средний риск заражения в этой популяции. Скорее всего, эту задачу, как и многие социально-эпидемиологические, можно ассимилировать к давно и хорошо проработанным статистическим методам. Прежде всего, нужно пойти на некоторые упрощения нашей исследуемой модели. Для этого придется сделать ряд предположений о предметной области.

Мы будем рассматривать имеющиеся у нас данные как набор интервалов между эпизодами, в том числе неточно определенных (с интервальной оценкой длины) и ограниченных с одной стороны (как, например, последний интервал перед интервью). При этом мы будем «забывать», от каких именно респондентов пришла эта информация. Имеющаяся статистическая информация теперь представляет собой просто набор длин интервалов, причем для некоторых из них эта длина задана неточно, а для некоторых — просто ограничена сверху или снизу.

Такие входные данные идеально укладываются в схему так называемой *статистики выживаемости* (*survival statistics*) [7]. Суть этого раздела статистики состоит в следующем. Имеется некоторая популяция (будь то прооперированные пациенты или лампочки), и интересующая нас информация — это время до выхода особей популяции из строя (смерть или перегорание), после которого они в популяцию уже обратно не возвращаются.

Особенность статистики выживаемости, благодаря которой она так хорошо подходит для решения стоящих перед нами задач, состоит в том, что там часто приходится учитывать недоопределенные данные. Например, человек, прооперированный десять лет назад, к моменту проведения исследования еще жив. Очевидно, длина интервала его жизни больше десяти лет, но никакой другой границы мы на него установить не можем. Это в точности соответствует ситуации последнего интервала перед интервью: длина интервала ограничена только с одной стороны. Такие недоопределенные данные в статистике выживаемости называют *цензурированными* (*censored data, censoring*).

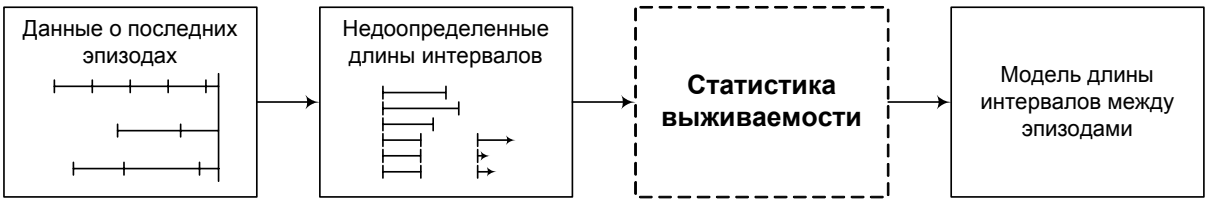

Рис. 1. Схема построения статистической модели.

В статистике выживаемости различают большое количество методов и подходов, направленных на единую цель — как можно более достоверно промоделировать поведение популяции с точки зрения выживаемости. Мы не будем вдаваться здесь в подробности математической статистики. В любом случае, результатом применения всех этих методов является функция, отражаю-

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

щая в том или ином виде (существуют как минимум четыре способа задать эту функцию) вероятность индивида выжить (или, наоборот, умереть) в течение заданного времени. Из такой функции можно будет стандартными методами теории вероятностей получить математическое ожидание количества эпизодов рискованного поведения («смертей») на любом заданном интервале. Схема предложенного алгоритма изображена на рис. 1.

Кроме того, следует отметить, что методы, описанные нами здесь, неприменимы для индивидуального подхода к каждому испытуемому: на первом же шаге мы «забываем», от кого какие интервалы поступили. Можно попытаться изобрести другие методы — например, продублировать одни и те же данные многократно, а затем применить вышеописанную статистику. Однако правомочность этих результатов нужно будет математически исследовать и подтверждать практикой — результатами исследования рискованного поведения.

8. Заключение

Основной характеристикой риска заразиться ВИЧ-инфекцией (а также скорости распространения ВИЧ-инфекции) является инсиденс-показатель. Его прямое измерение при помощи проведения когортного исследования занимает несколько лет и требует очень крупной исследовательской площадки. При этом суммы на исследование составляют миллионы долларов; также чрезвычайно высоки трудозатраты социальных работников, участвующих в исследовании.

Следует отметить, что в России систематическое проведение таких исследований пока не представляется реалистичным: ввиду недостаточного финансирования многие даже более значимые социальные проекты не получают должной поддержки. Говорить о каких-то серьезных достижениях в решении данной, новой для системы здравоохранения нашей страны, проблемы преждевременно. Ряд исследований, которые у нас проводятся, являются лишь аналогами зарубежных исследований по данной тематике [3].

В этой ситуации встает вопрос о применении косвенных измерений, в частности, подхода, который опирается на оценку риска заразиться по частоте участия респондента в различных видах рискованного поведения. Предполагается, что частота участия определяется по самоотчету респондента — по его ответам на вопросы специально разработанного опросника.

На самом деле мы рассматриваем не независимые виды поведения. Совершенно очевидно, что если человек употребил сильнодействующий внутривенный наркотик, то вероятность, что он будет использовать презерватив при половом контакте, сильно уменьшается, и соответственно увеличивается риск заразиться половым путем, а с ним — общий риск заражения. На самом деле это может являться еще одним очень серьезным направлением деятельности по данной тематике. Один из возможных вариантов — это использование алгебраических байесовских сетей для оценок зависимостей между различными видами поведения.

Ключевой параметр, на определение которого должен быть нацелен опрос, — это число эпизодов рискованного поведения, в которых респондент принял участие в течение какого-то заданного временно̀го интервала. Прямые вопросы и вопросы, ответ на которые представляется в виде Likert-шкалы, доставляют данные, непригодные для количественной оценки вероятности заразиться.

Труды СПИИРАН. 2006. Вып. 3. Т. 2. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 2. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Мы предлагаем рассмотреть возможность обработки ответов респондента на вопросы о небольшом числе последних эпизодов. Проведенное пилотное исследование показало, что наркопотребители вполне способны ответить на вопросы о трех последних эпизодах. В связи с этим в настоящей статье была представлена постановка задачи на адаптацию существующих классических математических методов для обработки достаточно «бедных» и нечетких данных, собираемых при опросе респондентов об их рискованном поведении.

Исследования, результаты которых изложены в настоящей статье, частично проводились в рамках государственного контракта 02.442.11.7489.

Литература

1. Доклад ЮНЕСКО об эпидемии СПИД в России [Электронный ресурс] // <http://www.aids.ru/aids/unaidsreport.shtml> (по состоянию на 10.05.2006).
2. *Тулупьев А. Л.* и др. Ответы наркопотребителей о последних эпизодах рискованного поведения // Развитие специальной (коррекционной) психологии в изменяющейся России: Материалы научно-практической конференции «Ананьевские чтения — 2005» / Под ред. Л. А. Цветковой, Л. М. Шипицыной. СПб.: Изд-во СПбГУ, 2005. С. 474–475.
3. *Крупицкий Е. М.* и др. Двойное слепое рандомизированное плацебоконтролируемое исследование эффективности Налтрексона для стабилизации ремиссий больных героиновой зависимостью // Ученые записки СПбГМУ им. академика И. П. Павлова. 2003. Т. X, № 2. С. 23–30.
4. *Микони С. В*. Теория и практика рационального выбора. М.: Маршрут, 2004. 462 с.
5. *Bell D. C, Trevino R. A.* Modeling HIV Risk [Epidemiology] // JAIDS. 1999. Vol. 22(3), November 01, 1999. P. 280–287.
6. Доклад ЮНЕСКО о глобальной эпидемии СПИДа [Электронный ресурс] // <http://www.unaids.org/bangkok2004/GAR2004_html_ru/GAR2004_02_ru.htm#TopOfPage> (по состоянию на 10.05.2006).
7. *Klein J. P., Moeschberger M. L.* Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer, 1997. 560 p.
8. Coverage of selected services for HIV/AIDS prevention and care in low and middle income countries in 2003 / USAID, UNAIDS, WHO, UNICEF, POLICY Project. Washington: POLICY Project, 2004. 84 p.
9. Доклад о глобальной эпидемии СПИДа — 2004 г.: Исполнительное резюме [Электронный ресурс] // <http://www.unaids.org/bangkok2004/GAR2004_ru_html/-ExecSumm_ru/ExecSumm_ru_01.htm#P50_13276> (по состоянию на 10.05.2006).
10. Финансирование мер по борьбе со СПИД [Электронный ресурс] // <http://www.unaids.org/bangkok2004/GAR2004_html_ru/GAR2004_10_ru.htm#P1227_268579> (по состоянию на 10.05.2006).
11. Основные понятия проблемы ВИЧ-инфекции [Электронный ресурс] // <http://www.infospid.ru/index.php?cat=saaa> (по состоянию на 10.05.2006).
12. *Rothman K. J.* Epidemiology: An Introduction. Oxford etc.: Oxford University Press, 2002. 223 p.