ОЦЕНКА ИНТЕНСИВНОСТИ ПОВЕДЕНИЯ РЕСПОНДЕНТА В УСЛОВИЯХ ИНФОРМАЦИОННОГО ДЕФИЦИТА

А. Е. Пащенко[1], А. Л. Тулупьев[2], Т. В. Тулупьева[3]

[1,2,3]Санкт-Петербургский институт информатики и автоматизации РАН

[1,2,3]СПИИРАН, 14-я линия ВО, д. 39, Санкт-Петербург, 199178

[1]<aep@iias.spb.su>, [2]<alt@iias.spb.su>, [3]<tvt@iias.spb.su>

<http://tulupyev.spb.ru>

УДК 311.2 + 616-036.22

Пащенко А. Е., Тулупьев А. Л., Тулупьева Т. В. **Оценка интенсивности поведения респондента в условиях информационного дефицита** // Труды СПИИРАН. Вып. 7. — СПб.: Наука, 2008.

Аннотация. *В статье описан подход к оценке интенсивности поведения, математической моделью которого является пуассоновский случайный процесс, моделирующий это поведение как серию эпизодов. Оценка интенсивности формируется в условиях дефицита информации, а именно на основе сведений о небольшом числе последних эпизодов, причем эти сведения представлены неточными высказываниями респондента на естественном языке.* — Библ. 12 назв.

UDC 311.2 + 616-036.22

Paschenko A. E., Tulupyev A. L., Tulupyeva T. V. **Respondents Behaviour Intensity Estimation under the Information Deficiency** // SPIIRAS Proceedings. Issue 7. — SPb.: Nauka, 2008.

Abstract. *In article the approach to an estimation of intensity of the behavior which mathematical model is Poisson the casual process modeling this behavior as a series of episodes is described. The intensity estimation is formed in the conditions of deficiency of the information, namely on the basis of data on a small number of last episodes, and these data are presented by inexact statements of the respondent in a natural language.* — Bibl. 12 items.

1. Введение

В ряде отраслей научных исследований стоит задача оценки интенсивности поведения респондентов по неполным и неточным исходным данным. Например, в случае сахарного диабета ключевым показателем является частота отклонения пациента от диеты. Строгое соблюдение низкоуглеводородной диеты — существенное условие для замедления развития сахарного диабета [1]. Лечащему врачу необходимо знать, насколько эффективными оказались его рекомендации по поводу соблюдения диеты; такие сведения помогут ему понять, какие действия необходимо предпринять для более эффективного лечения. Оценка эффективности рекомендаций, а также выбор тактики дальнейшего лечения пациента строятся (явно или неявно) исходя из степени интенсивности отклонения от диеты.

В последние годы в России неуклонно распространяется эпидемия инфекции, вызванной вирусом иммунодефицита человека (ВИЧ). В силу особенностей ВИЧ-инфекции (неизлечимость, высокая социальная и социально-эпидемиологическая опасность) она превратилась в одну из наиболее серьезных проблем здравоохранения в Российской Федерации. Вакцины против ВИЧ пока не существует, поэтому единственным методом профилактики распространения эпидемии ВИЧ/СПИДа во всем мире являются превентивные программы, направленные в первую очередь на модификацию поведения людей. Эффективность таких программ необходимо оценивать. Ключевым параметром

при оценке рисков передачи ВИЧ-инфекции различными видами поведения является его интенсивность. Оценка эффективности превентивных программ базируется на том, как измеряются соответствующие риски.

Важным достижением, повлиявшим на исследования, посвящённые изучению жизни ВИЧ-инфицированных людей и продлению их жизни, явилась разработка высокоактивной антиретровирусной терапии (ВААРТ). Под терапией подразумевается ежедневный прием препаратов по строгому графику (расписанию). При правильном лечении ВААРТ позволяет остановить размножение вируса. Для того чтобы убедиться в этом, делается тест на вирусную нагрузку (ВН) — определяется, сколько вирусов содержится в одном миллилитре крови. ВН определяет, насколько быстро будет снижаться уровень иммунитета, как быстро вирус будет уничтожать CD4 клетки. В тяжелых случаях ВН может доходить до миллиона копий вируса в миллилитре [2]. После того как размножение вируса прекращается, у человека постепенно восстанавливается иммунитет. ВААРТ-терапия продлевает жизнь человека. Есть люди, принимающие терапию и благодаря этому живущие с ВИЧ более 25 лет.

Для того чтобы АРВ-терапия имела положительный результат, пациент должен строго соблюдать режим лечения. Из всех назначенных таблеток нужно выпить вовремя как минимум 95 %. То есть для большинства схем лечения допустимым является пропуск не более трех дозировок лекарств в месяц [3].

Приверженность к лечению является одним из наиболее важных условий, позволяющих достичь высокой эффективности лечения ВИЧ-инфекции. Низкий уровень приверженности к ВААРТ может привести не только к снижению эффективности лечения, но и к возникновению устойчивых форм ВИЧ. В результате снизится количество приемлемых вариантов комбинации препаратов для лечения [3].

Наилучший результат ВААРТ наблюдается при 100 %-ной приверженности к лечению. Уровни ниже 95 % (пропуск или запаздывание с приемом каждой двадцатой дозы) уже могут привести к ослаблению подавления вируса и более медленному росту концентрации клеток CD4. Несмотря на то что большинство людей понимает, что своевременный и обязательный прием препаратов должен стать привычной частью их жизни, многим не удается достичь таких высоких показателей приверженности к лечению. На практике оказывается, что это достаточно сложно при приеме ВААРТ-терапии; при изучении этой проблемы ключевой является оценка интенсивности отклонения от графика приема препаратов. Если отклонение больше определенного «порогового» значения, то необходимо принимать срочные меры по коррекции поведения пациента.

Выходя за рамки рискованного поведения, предложенный ниже метод можно использовать для оценки потребления тех или иных товаров или продуктов. В частности, можно выделить группы потребителей, существенно различающиеся интенсивностью потребления продуктов, товаров или услуг. При наличии таких результатов маркетинговые усилия можно сосредоточить на тех группах, которые многочисленны, но товар потребляют неинтенсивно. Такая стратегия может привести к существенному увеличению объема продаж.

Следует отметить, что необходимые данные об интенсивности потребления невозможно получить из анализа продаж, то есть недостаточно изучить «чеки» — данные о состоявшихся продажах. Это позволит принять во внимание лишь те группы, которые и так уже покупают данный товар. Не исключено, что в таком случае из анализа выпадут многочисленные потенциальные потребите-

Труды СПИИРАН. 2008. Вып. 7. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 7. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

ли, которые ни разу еще не употребляли интересующие маркетологов товары или услуги.

Под интенсивностью понимается число эпизодов поведения рассматриваемого вида в определенный промежуток времени. Существует несколько подходов, которые позволяют регистрировать эпизоды поведения различного вида. В частности, «дневниковый метод» подразумевает запись всех действий респондента в течение дня, после чего полученные данные, как правило, накапливавшиеся несколько месяцев, поступают на обработку эксперту, который подсчитывает число эпизодов поведения определенного вида за данный период. Однако такой вид исследований достаточно дорог, его сложно организовать и долго выполнять. Поэтому встает задача оценки интенсивности рискованного поведения респондента по его «одномоментному» самоотчету, то есть по ответам на блок вопросов или по результатам проведения интервью. Заметим, что подобные опросы опираются на информацию, хранящуюся в памяти респондента, и, естественно, чем глубже ретроспектива, тем труднее респондентам отвечать на вопросы и тем больше они делают ошибок припоминания.

На данный момент разработаны и применяются в опросах респондентов два подхода к оцениванию интенсивности поведения, каждый из которых имеет недостатки. Первый подход — 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. Попытка ответа даже самому себе на следующие, казалось бы, простые вопросы, касающиеся повседневных действий: «Сколько раз за последние три месяца я пил чай с сахаром?» или «Сколько фруктов я съел за последние полгода?», вызывает сложность. Получается, что для достоверного ответа на данные вопросы респондент должен оценить среднее число эпизодов в день, понять, какие параметры влияют на данную величину, например день недели, температура воздуха, время суток и т. д., вспомнить, не было ли за интересующий период различных стилей поведения (закончился чай — месяц не пил), и в завершение дать итоговую взвешенную оценку числа эпизодов. Становится понятным, насколько сложно отвечать на данные вопросы, в особенности респондентам из маргинальных социальных групп.

Второй метод — лайкерт-шкалы [5] — опросники, в которых используются качественные, а не количественные варианты: «Никогда», «Редко», «Иногда», «Часто», «Всегда», и подобные им возможности для ответа. Вопрос ставится легко, ответы на них 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 необязательно совпадает с расстоянием между «Редко» и «Никогда». На практике шкалы арифметизируют, но за этой арифметизацией не стоит никакой сколько-нибудь правдоподобной гипотезы; ситуацию с риском получающиеся расчеты не характеризуют вообще никак. Таким образом, возникает потребность в более адекватных источниках сведений о поведении конкретного вида и методиках их обработки, которые сделают возможной более обоснованную оценку числа эпизодов или целевых показателей.

Одной из возможных альтернатив по использованию лайкерт-шкал представляется опрос респондента об одном или нескольких последних эпизодах его поведения. Такой опрос позволяет судить о непрерывных количественных

Труды СПИИРАН. 2008. Вып. 7. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 7. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

величинах — интервалах между эпизодами, а также об интервале между временем опроса и последним эпизодом. Уже сами упомянутые интервалы могут оказаться более удобными и достоверными косвенными показателями риска по сравнению с порядковыми шкалами вида «Никогда–Редко–Иногда–Часто–Всегда». Чем меньше интервалы, тем более высокую степень риска мы можем ожидать. Интервалы можно сравнивать между собой.

Практическая апробация метода косвенных измерений, основанного на данных о последних эпизодах рискованного поведения, была осуществлена Санкт-Петербургским институтом информатики и автоматизации РАН на базе на базе СПб ГУЗ «Центр по профилактике и борьбе со СПИДом и инфекционными заболеваниями» (СПИД-Центр) [6,10].

В первом исследовании приняли участие 160 ВИЧ-инфицированных. Было опрошено 73 женщины, что составляет 45.6% от общего числа испытуемых и 87 мужчин, что составляет 54.4% в возрасте от 18 лет, которые находились на различных стадиях развития ВИЧ-инфекции [10].

Во втором исследовании принял участие 301 ВИЧ-инфицированный пациент, из них 135 (44.9%) мужчин и 166 (55.1%) женщин в возрасте от 18 лет, которые находились на различных стадиях развития ВИЧ-инфекции и получили эту инфекцию разными путями [4, 6].

С каждым из пациентов было проведено персональное интервью, и каждый из участников прошел психологическое тестирование [4, 6, 7, 10].

В частности, в ходе исследований регистрировались ответы о рискованном поведении, связанном: а) с отклонением от графика приема препаратов антиретровирусной терапии; б) с употреблением алкоголя; в) с употреблением внутривенных наркотиков. Задавались вопросы о последних трех эпизодах рискованного поведения, максимальном, минимальном и обычном интервалах между эпизодами. Ответы фиксировались в тех словесных формулировках, которые использовал сам респондент.

По результатам исследований можно утверждать, что респонденты без затруднений отвечают на сформулированные вопросы, интервьюеры также фиксировали дополнительные сведения из высказываний респондентов, которые полезны для формирования более точной оценки риска передачи ВИЧ-инфекции: минимальный, максимальный и обычный интервалы между эпизодами рискованного поведения.

Было установлено, что ответы респондентов достаточно стереотипны, что позволило разработать классификацию формулировок ответов. Выделены два относительно независимых атрибута, характеризующие совокупность формулировок ответов, и шесть их классов, требующих различных подходов и процедур для последующей обработки [6, 9].

Целью работы является описание методики оценки интенсивности деятельности респондента на основе его неточных и недоопределенных ответов об его участии в этой деятельности или в данном поведение, а также описание математических моделей, лежащих в основе процедуры оценивания.

2. Классификация ответов

В предыдущих статьях [6, 7] была представлена частная классификация ответов респондентов, связанная с их ВИЧ-рискованным поведением. В данной статье представлена обобщенная модель классификации ответов на естественном языке. Представленная ниже схема оценки ответов покрывает все воз-

Труды СПИИРАН. 2008. Вып. 7. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 7. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

можные варианты ответов об эпизодах рискованного поведения. Любой ответ респондентов можно свести к одной из предложенных формальных схем.

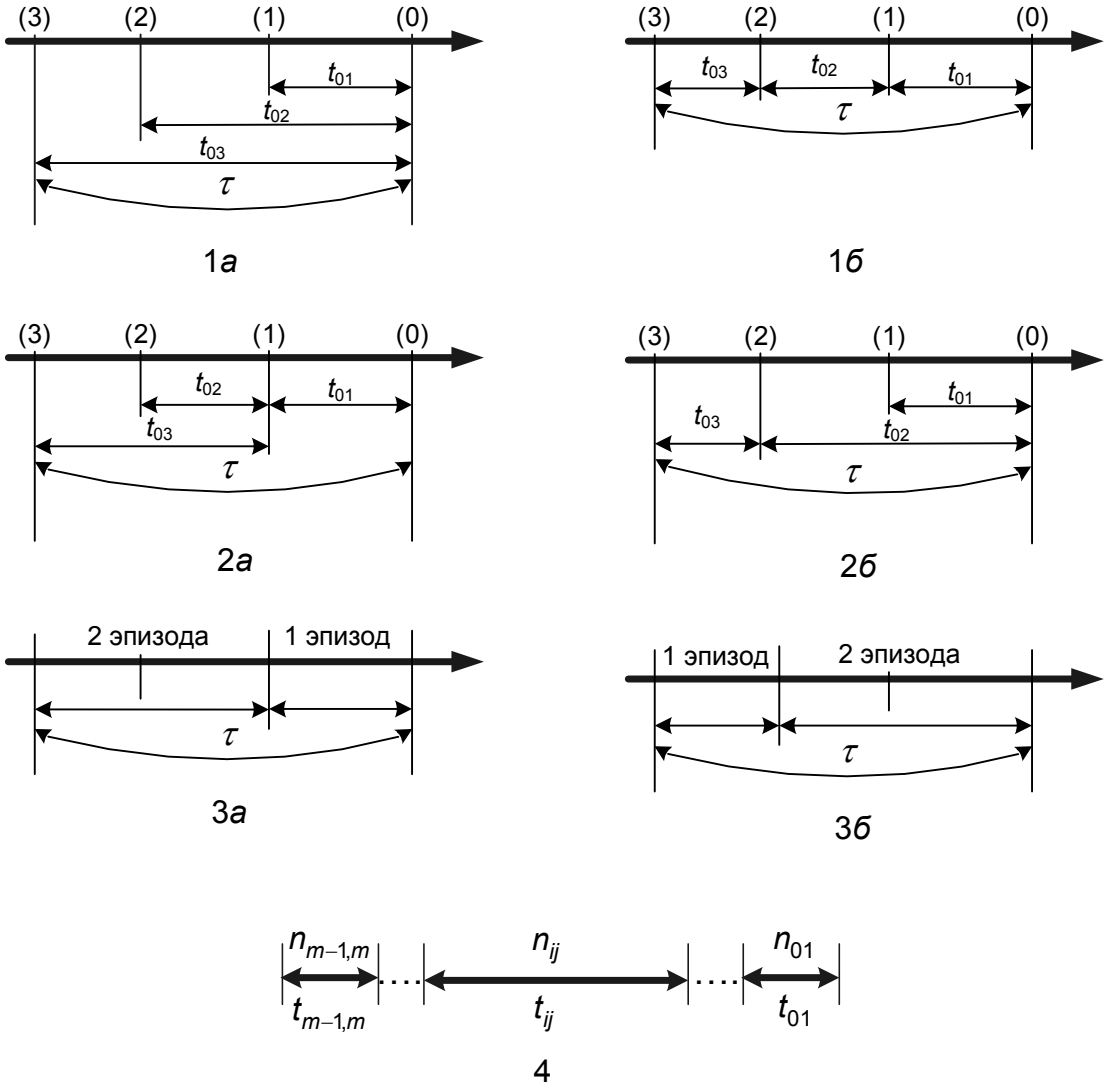

Рис. 1. Обобщенная классификация ответов.

На рис. 1 схематично представлены формализованные варианты ответов респондентов, где (0) — момент интервью; (1), (2), (3) — моменты на оси времени, когда произошел последний, предпоследний и пред-предпоследний эпизод поведения; t_{01}, t_{12}, t_{23} — длины временных интервалов соответственно между моментом интервью и последним эпизодом, последним и вторым эпизодом, вторым и третьим эпизодом поведения в прошлом; τ — временной промежуток, за который произошли эпизоды.

На рис. 1 (1*а*) представлен вариант ответов, когда респондент указывает временной интервал между моментом интервью и каждым эпизодом, например: «вчера, позавчера, поза-позавчера».

На рис. 1 (1*б*) представлен вариант ответов, когда респондент указывает эпизоды рискованного поведения начиная с предпоследнего, отсчитывая их от момента предыдущего эпизода, например: «вчера, за неделю до этого, за неделю до этого».

Труды СПИИРАН. 2008. Вып. 7. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 7. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

На рис. 1 (2*а*, 2*б*) представлены варианты ответов, являющиеся комбинацией предыдущих двух классов, например: «вчера, позавчера, еще за день до этого».

На рис. 1 (3*а* и 3*б*) схематично изображены возможные последовательности ответов респондентов, при этом дается одна оценка для отдельной последовательности, а общая оценка является суммой оценок последовательностей.

На рис. 1 (4) предложена обобщенная схема представления ответов респондентов, в рамках которой эксперт должен действовать. То есть рис. 1 показывает, что эксперт имеет дело лишь с объединенными в последовательности эпизодами поведения, для которых (последовательностей ответов) он дает оценку, а итоговая оценка складывается из суммы оценок таких последовательностей.

3. Оценка методом максимального правдоподобия

Цель наших дальнейших исследований — определить или оценить величину параметра λ, характеризующего интенсивность участия респондента в поведении определенного вида, которое описывается пуассоновским случайным процессом. Получив оценку параметра λ, можно посчитать вероятность того, что в интервале $[t_0, t_0 + t]$ произойдут k событий:

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

Оценка может быть получена методом максимального правдоподобия. Пусть дано ν — число последовательных эпизодов от момента интервью, которые вспомнил респондент, а $\tau = t_{0\nu}$ — тот период времени, за который эти эпизоды произошли. Применим метод максимального правдоподобия к основному уравнению пуассоновского процесса при вышеуказанных данных, чтобы найти соответствующую оценку интенсивности λ:

$$g(\lambda) = \frac{(\tau\lambda)^\nu}{\nu!} e^{-\tau\lambda},$$

$$h(\lambda) = \ln g(\lambda) = \nu \ln\lambda + \nu \ln\tau - \ln\nu! - \tau\lambda,$$

$$\frac{dh(\lambda)}{d\lambda} = \frac{\nu}{\lambda} - \tau,$$

$$\frac{dh(\lambda)}{d\lambda} = 0 \Rightarrow \lambda = \frac{\nu}{\tau}.$$

Заметим, что выкладки привели к тому же самому заключению, что и в [7].

Как правило, исходя из ответов респондентов удается дать численную оценку числу произошедших эпизодов между моментом интервью и наиболее отдаленным эпизодом включительно. Если, в частности, респондент ответил на все вопросы о последних трех эпизодах, то $\nu = 3$.

В силу существенной неопределенности высказываний на естественном языке получить точную численную оценку τ (в рассматриваемом случае τ — это величина временного интервала между моментом интервью и самым отдаленным от него эпизодом включительно) затруднительно или даже невозможно. Однако ее можно рассмотреть как случайную величину, построенную над дру-

Труды СПИИРАН. 2008. Вып. 7. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 7. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

гими случайными величинами. Рассмотрим особенности процесса построения такой случайной величины.

4. Рандомизация

4.1. Оценка временного интервала

Интервалы t_{ij} (из которых складывается τ) измерены неточно, они характеризуются существенной недоопределенностью, связанной с тем, что результаты «измерения» зафиксированы на естественном языке, терминами повседневной речи — с привычной и приемлемой для бытовых нужд строгостью и точностью.

Разберем оценку длины временного интервала t_{ij}. Во-первых, отметим, что респонденты используют в своих высказываниях преимущественно следующие единицы измерения: часы, дни, недели, месяцы, полгода, годы.

Во-вторых, обратим внимание, что использованная единица измерения несет в себе информацию о точности измерения. Поясним это на примере двух высказываний: «семь дней назад» и «неделю назад». Когда респондент использует формулировку «семь дней назад», это свидетельствует о высокой «надежности» припоминания и его уверенности в том, что событие произошло ровно семь дней назад. Обобщая сказанное выше, можно сделать предположение, что если событие произошло за период более длинный, чем количество единиц измерения (например, *семь* дней), которых достаточно для построения высказывания, используя другую «следующую» языковую конструкцию (*одну* неделю назад), то можно говорить о повышенной степени точности, а также, вероятно, об экстраординарности произошедшего события, которая позволила запомнить его гораздо лучше и выделить из ряда других повседневных событий. Когда респондент использует формулировку «неделя назад», он априорно снижает точность высказывания. Неделя назад — это и шесть дней назад и восемь. При этом если речь идет о высказывании, например, «семь недель», мы опять имеем повышенную точность. Это опять может свидетельствовать об экстраординарности события, обычно в таких случаях употребляется конструкция «два месяца назад».

Можно ввести определение, что под «нормальной точностью» понимается точность единиц измерения, а также кратные ей единицы, если они не могут быть заменены такой языковой конструкцией, как например «семь дней» $\rightarrow$ «неделя», «четыре недели» $\rightarrow$ «месяц» и так далее. Под «повышенной точностью» будем понимать, высказывания, которые являются временными промежутками, которые могли быть заменены языковой конструкцией следующего порядка, например «двадцать три дня назад», «14 месяцев назад».

То есть для человеческого восприятия «непрерывного времени» уже выработаны внутренние механизмы перехода от одной точности высказывания (более высокой) к другой (более грубой) в зависимости от отдаленности события в прошлом. При этом ряд высказываний могут обладать более высокой точностью.

При проведении исследования [7] установлено, что ряд респондентов могли называть точную дату употребления ими наркотиков в десятилетней ретроспективе.

Труды СПИИРАН. 2008. Вып. 7. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 7. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Соответственно для грубой оценки интервала t_{ij} предлагается использовать не только скалярную величину $L(t_{ij})$, но также и нижнюю границу $L^{-}(t_{ij})$ вместе с верхней границей $L^{+}(t_{ij})$: $L^{-}(t_{ij}) \le L(t_{ij}) \le L^{+}(t_{ij})$.

4.2. Неточность высказываний респондента

1. Неточность высказываний респондента требует также введения недетерминированной части оценки. Ее основу составляет случайная величина ξ, которая отвечает ряду требований:

а) $\mathbf{M}\xi = 0$;

б) плотность распределения f_{ξ} указанной случайной величины ξ симметрична относительно нуля: $f(t) = f(-t)$;

в) плотность распределения f достигает своего максимума в нуле, до нуля она монотонно возрастает, после — монотонно убывает (требование строгой монотонности не предъявляется);

г) носитель плотности распределения f конечен.

Заметим, что требование б) может и не выполнятся, если респонденты имеют тенденцию либо преувеличивать, либо преуменьшать оценки.

Для практических приложений можно рассмотреть еще более строгое требование, ограничивающее величину стандартного отклонения (квадратный корень из дисперсии случайной величины): например $2\sigma \le \frac{1}{4}$. Кроме того, разброс значений случайной величины можно ограничить указанием верхней и нижней границ, а также межквартильного размаха.

4.3. Единицы измерения

В высказываниях респондентов используются следующие единицы измерения времени: δ_1 — часы, δ_2 — дни, δ_3 — недели, δ_4 — месяцы, δ_5 — полгода, δ_6 — года.

Таблица 1

Приведенные к дням единицы измерения

Единица измерения	Единица измерения, приведенная к дням	Значения погрешности, в днях	Значение погрешности при d=1/4 (десят.)	Значение погрешности при d=1/4 (обыкн.)
Часы	1/24	±1/24 *d*	±0.01	±1/96
Дни	1	±1 *d*	±0.25	±1/4
Недели	7	±7 *d*	±1.75	±7/4
Месяцы	30	±30 *d*	±7.50	±7½
Полгода	183	±183 *d*	±45.75	±45¾
Годы	366	±366 *d*	±91.50	±91½

4.4. Интерпретация высказываний

Высказывания на естественном языке позволяют один и тот же временной интервал представить различными языковыми конструкциями. Перечислим основные встречающиеся тождественные высказывания и приведем их к стандартному виду, к дням:

Труды СПИИРАН. 2008. Вып. 7. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 7. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Таблица 2

Приведенные к дням единицы измерения

Исходное	Стандартизированное
Сегодня	1/4 дня
Вчера	Один день
Позавчера	Два дня
Поза-позавчера	Три дня
На прошлой неделе	В зависимости от дня проведения интервью
В прошлом году	В зависимости от месяца проведения интервью

4.5. Примеры интерпретации ответов

Когда высказывание имеет вид «вчера, позавчера, позапозавчера» и интенсивность характеризуется «случаями в день», то длина интервала τ будет равна $\tau = x\delta_2 + \xi\delta_2 = 3\times 1 + \xi = 3 + \xi$. Если предположить, что $2\sigma \le \frac{1}{4}\delta_2$, тогда можно перейти к интервальной оценке τ: $3 - \frac{1}{4} \le \tau \le 3 + \frac{1}{4}$.

Когда высказывание имеет вид «неделю назад, еще неделю назад, еще неделю назад» и интенсивность характеризуется «случаями в день», то

$$\tau = (x_1\delta_3 + x_2\delta_3 + x_3\delta_3) + (\xi_1\delta_3 + \xi_2\delta_3 + \xi_3\delta_3) = (7+7+7) + (7\xi_1 + 7\xi_2 + 7\xi_3) = 21 + 7\sum_{i=1}^{3}\xi_i.$$

Если предположить, что $2\sigma \le \frac{1}{4}\delta_2$, тогда можно перейти к интервальной оценке τ:

$$21 - \frac{21}{4} \le \tau \le 21 + \frac{21}{4},\ 21 - 5\frac{1}{4} \le \tau \le 21 + 5\frac{1}{4}.$$

Если предположим, что опрос проводился 1 июня, а ответы имели вид «8 марта, 23 февраля, 1 января», то $\tau = 4\times 30.5 + \xi = 122 + \xi$. Если предположить, что $2\sigma \le \frac{1}{4}\delta_2$, тогда можно перейти к интервальной оценке τ:

$$122 - \frac{1}{4} \le \tau \le 122 + \frac{1}{4}.$$

Когда высказывание имеет вид «вчера, позавчера, еще за неделю до этого» и интенсивность характеризуется «случаями в день», то

$$\tau = t_{02} + t_{23} = 2 + \xi + 7 + 7\xi = 9 + 8\xi.$$

Если предположить, что $2\sigma \le \frac{1}{4}\delta_2$, тогда можно перейти к интервальной оценке τ: $9 - 2 \le \tau \le 9 + 2$.

4.5. Арифметизация неопределенности

Отрезок $\left[L^-, L^+\right]$ дает область всех возможных значений $\hat{L}(t_{ij})$ — длины временного интервала t_{ij}; $L(t_{ij})$ является скалярной (точечной) оценкой длины указанного интервала.

Труды СПИИРАН. 2008. Вып. 7. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 7. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

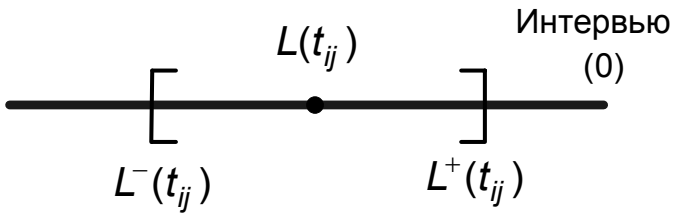

Рис. 2. Отрезок $\left[L^-, L^+\right]$ на оси времени.

Заметим, что любая точка из интервала $\left[L^-, L^+\right]$ *возможна* в качестве значения $\hat{L}(t_{ij})$; однако это не означает, что точки из этого интервала «равноправны» или равновероятны в качестве значения $\hat{L}(t_{ij})$.

Сведения о такого рода отношениях между допустимыми значениями можно задать с помощью их распределения вероятностей. В таком случае $\hat{L}(t_{ij})$ будет представлять собой случайную величину, что позволит нам использовать вероятностные методы для представления и обработки неопределенности исходных данных [12].

В зависимости от наших предположений о характере ответов респондента, для задания случайной величины $\hat{L}(t_{ij})$ мы можем использовать *непрерывные* распределения: равномерное, треугольное, трапециевидное, β-распределение — или *дискретные* распределения: равномерное, биномиальное (после определенной адаптации), другие дискретные индуцированные распределения, представленные ранее перечисленными непрерывными.

Рассмотрим более подробно случай двух первых дискретных распределений.

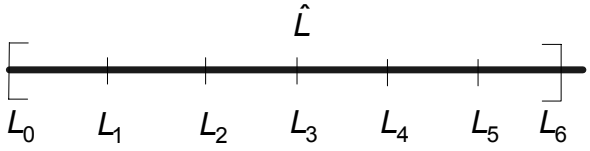

Рис. 3. Разбиение отрезка $\hat{L}$.

Область возможных значений разбивается на N отрезков одинаковой длины. Границы отрезков нумеруются от 0 до n; на рис. 3 — для $n = 6$.

В случае равномерного дискретного распределения все значения равновероятны, то есть

$$\forall i : 0 \le i \le n \quad p(L_i) = \frac{1}{n+1}.$$

В случае биномиального распределения с заданной вероятностью успеха π

$$\forall i : 0 \le i \le n \quad p(L) = C_n^i \pi^i (1-\pi)^i, \pi \in [0,1].$$

Заметим, что в случае дискретного равномерного распределения, так и в случае биномиального распределения при $\pi = \frac{1}{2}$, справедливо $\mathbf{E}\hat{L}(t_{ij}) = L(t_{ij})$.

Отклонение параметра биномиального распределения от 1/2 позволяет моделировать ситуацию, когда респондент дает оценку со смещением в большую или меньшую сторону.

Труды СПИИРАН. 2008. Вып. 7. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 7. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Из случайных величин $\hat{L}(t_{ij})$ строится случайная величина $\hat{L}(\tau)$ — длина временного интервала от момента интервью до самого отдаленного от него эпизода.

В зависимости от конкретного класса ответов эта величина будет либо равна одной из величин $\hat{L}(t_{ij})$, либо будет являться суммой каких-то из них.

Рассмотрим наиболее сложный для сведений случай из трех эпизодов:

$$\hat{L}(\tau) = \hat{L}(t_{01}) + \hat{L}(t_{12}) + \hat{L}(t_{23}).$$

Множество возможных значений $\hat{L}(\tau)$ будет состоять из всех возможных сумм вида

$$L_{ijk}(\tau) = L_i(t_{01}) + L_j(t_{12}) + L_k(t_{23}),$$
$$0 \le i, j, k \le n.$$

Причем каждой такой сумме будет составлена вероятность $p(L_i) \times p(L_j) \times p(L_k)$.

Чтобы получить распределение случайной величины $\hat{L}(\tau)$, необходимо сгруппировать одинаковые по величине суммы; вероятность соответствующего значения будет складываться из вероятностей этих сумм:

$$\hat{L}_{i,j,k}(\tau) = L,$$
$$p(L) = \sum_{L_i + L_j + L_k = L} p(L_i) \times p(L_j) \times p(L_k).$$

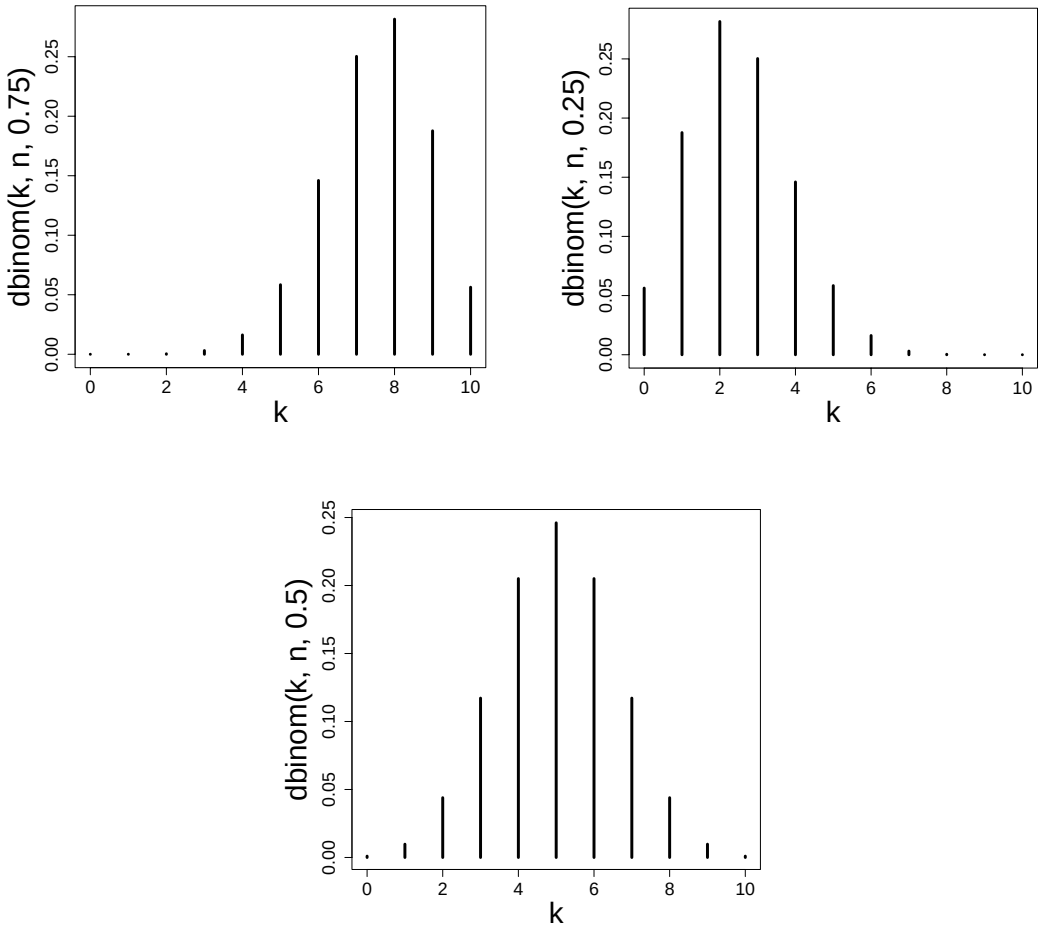

Рис. 4. Биномиальное распределение при $n = 10k$, $\pi = 0.75$, $\pi = 0.25$, $\pi = 0.50$.

Труды СПИИРАН. 2008. Вып. 7. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 7. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

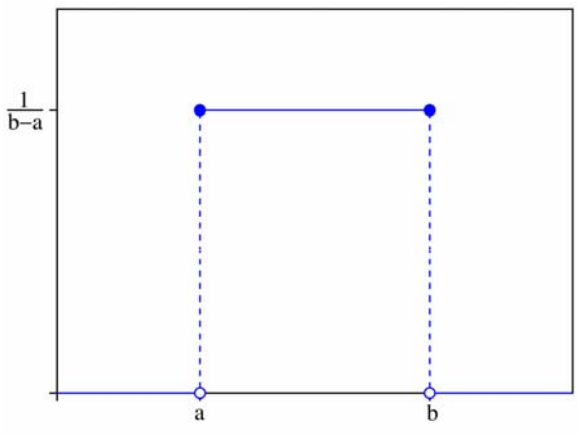

Рис. 5. Равномерное распределение.

Мы можем от рандомизированной оценки длины интервала $\hat{L}(\tau)$ перейти к случайной величине $\hat{\lambda}$, характеризующей интенсивность поведения:

$$\lambda_{ijk} = \frac{\upsilon}{L_i + L_j + L_k}.$$

Заданные числа формируют множество значений, и вследствие этого можем перейти к случайной величине $\hat{\lambda}$.

Для данной величины можно вычислять математическе ожидание, дисперсию и стандартное отклонение, а также межквартальный размах. Следует оговориться, что для некоторых классов ответов респондентов все значения сочетания L_i, L_j, L_k возможны, например в случае ответов «вчера, позавчера, позавчера». При таких условиях невозможным сочетаниям приписывается вероятность 0; возможным сочетаниям «вероятности» приписываются согласно приведенным формулам, а потом нормируются.

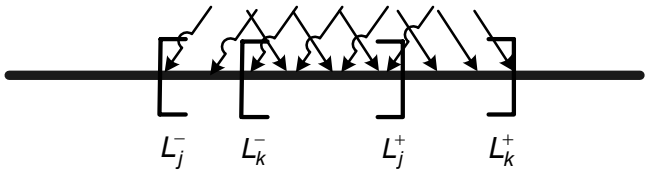

Рис. 6. Невозможные сочетания ответов.

Далее предполагаем, что после рандомизации нам удалось сформировать дискретную случайную величину $\hat{\lambda}$. Ее возможным значениям $\{\lambda_m\}_{m=0}^{M}$ приписаны вероятности $p(\lambda_m) = p_m$, отвечающие следующим требованиям:

$$\forall m \quad p_m \geq 0,$$

$$\sum_{m=0}^{M} p_m = 1.$$

Труды СПИИРАН. 2008. Вып. 7. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 7. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

5. Идентификация интенсивности по экстремальным интервалам

Особый интерес представляют ответы респондентов, содержащие сведения о максимальном и минимальном интервале между эпизодами рискованного поведения за заданный период времени. Во время проведенных опросов выяснилось, что, как правило, респондент без напряжения может вспомнить и указать такие «экстремальные» значения. В силу доступности таких данных хотелось бы иметь способ оценить на их основе интенсивности рискованного поведения, в качестве модели которого выступает пуассоновский процесс.

Дадим формальную постановку задачи.

Пусть моделью рискованного поведения выступает пуассоновский процесс с основным уравнением

$$\Pr\left(\Delta t,k,\lambda\right)=\frac{\left(\lambda\Delta t\right)^k}{k!}e^{-\lambda\Delta t},$$

где Δt — промежуток времени наблюдения за поведением респондента; k — число эпизодов рискованного поведения, случившихся в этот промежуток; λ — интенсивность рискованного поведения; $\Pr\left(\Delta t,k,\lambda\right)$ — вероятность того, что за промежуток времени наблюдения при рискованном поведении с интенсивностью λ случится ровно k эпизодов указанного поведения.

Заметим, что для такого пуассоновского процесса известна также плотность распределения T — длины временного интервала между двумя соседними эпизодами; это — экспоненциальная плотность, которая задается следующим уравнением:

$$p(T)=\lambda e^{-\lambda T},\quad T\geq 0,$$

а соответствующая функция распределения —

$$\Psi(T)=\int_0^T\lambda e^{-\lambda\tau}\mathrm{d}\tau=-e^{-\lambda\tau}\Big|_0^T=1-e^{-\lambda T}.$$

Пусть заданы $T_{\min}$ — минимальная длина временного интервала между двумя соседними эпизодами и $T_{\max}$ — максимальная длина временного интервала между двумя соседними эпизодами.

Вопрос состоит в том, как получить оценку интенсивности по имеющимся данным:

$$\hat{\lambda}=\hat{\lambda}\left(T_{\min},T_{\max}\right).$$

Можно предложить несколько подходов. В частности, было бы интересно изучить распределения экстремальных длин интервалов $T_{\min}$ и $T_{\max}$ между двумя соседними эпизодами в зависимости от времени наблюдения Δt и интенсивности процесса λ, а затем воспользоваться одной из модификаций метода максимального правдоподобия. Однако мы рассмотрим подход, который нацелен на максимизацию вероятности того, что длины интервалов между соседними эпизодами попадают в замкнутый промежуток $\left[T_{\min},T_{\max}\right]$. Это требование тоже представляется обоснованным в силу того, что вполне естественно ожидать, что длины всех таких интервалом должны заключаться между своими нижней и верхней границами, которые, как раз, и представления соответствующими значениями $T_{\min}$ и $T_{\max}$.

Труды СПИИРАН. 2008. Вып. 7. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 7. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Вычислим соответствующую вероятность как функцию от неизвестной интенсивности процесса λ.

$$f(\lambda) = p\left(T \in [T_{\min}, T_{\max}]\right) = \int_{T_{\min}}^{T_{\max}} \lambda e^{-\lambda\tau} d\tau = \Psi(T_{\max}) - \Psi(T_{\min}) = e^{-\lambda T_{\min}} - e^{-\lambda T_{\max}}.$$

Определим, при каком значении λ функция $f(\lambda)$ приобретает максимальное значение:

$$f'(\lambda) = -T_{\min} e^{-\lambda T_{\min}} + T_{\max} e^{-\lambda T_{\max}} = T_{\max} e^{-\lambda T_{\max}} - T_{\min} e^{-\lambda T_{\min}} = 0;$$

$$T_{\max} e^{-\lambda T_{\max}} = T_{\min} e^{-\lambda T_{\min}};$$

$$\ln T_{\max} - \lambda T_{\max} = \ln T_{\min} - \lambda T_{\min};$$

$$\ln T_{\max} - \ln T_{\min} = \lambda T_{\max} - \lambda T_{\min};$$

$$\ln T_{\max} - \ln T_{\min} = (T_{\max} - T_{\min})\lambda;$$

$$\lambda = \frac{\ln T_{\max} - \ln T_{\min}}{T_{\max} - T_{\min}}.$$

Полученная таким способом оценка интенсивности будет вычисляться по формуле:

$$\hat{\lambda} = \frac{\ln T_{\max} - \ln T_{\min}}{T_{\max} - T_{\min}} = \frac{\ln \dfrac{T_{\max}}{T_{\min}}}{T_{\max} - T_{\min}}.$$

Свойства этой оценки еще предстоит исследовать, равно как и адаптировать процесс ее вычислений к неточности и нечеткости данных, которые удается извлечь и высказываний респондента. Кроме того, потребуется изучить вопросы, связанные с согласованностью оценок, которые получаются по данным о последних трех эпизодах, об «обычном» интервале между эпизодами, а также по данным о максимальном и минимальном интервале времени между соседними эпизодами.

В данной работе мы закончим описание предлагаемого подхода, ограничившись лишь рассмотрением примера, в котором обрабатывается достаточно часто встречающийся ответ респондента. Пусть он ответил, что минимальный промежуток между приемом наркотиков был ровно сутки, а максимальный — неделя. Тогда, отбрасывая анализ неточностей и нечеткостей, получим $\hat{\lambda} = \dfrac{\ln 7 - \ln 1}{7 - 1} = \dfrac{\ln 7}{6} \approx 0.324$, что примерно соответствует «среднему» интервалу между соседними эпизодами в трое суток.

6. Приложение к оценке вероятностей (рисков) получения или передачи ВИЧ-инфекции

Развитие современной эпидемиологии требует разработки более дешевых и оперативных косвенных методик измерения инсиденс-показателя на основе ограниченного набора сведений.

D. C. Bell и R. A. Trevino разработали математическую модель оценки риска заражения ВИЧ в результате рискованного сексуального и инъекционного поведения [8]. Основные параметры модели, предложенной авторами, таковы:

$$\mathrm{Pr}_i = 1 - (1 - p_i)^{N_i},$$

Труды СПИИРАН. 2008. Вып. 7. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 7. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$$\Pr = 1 - \prod_{i=1}^{n} (1 - \Pr_i),$$

где n — число видов рискованного поведения; p_i — вероятность заразиться за отдельный эпизод i-го вида рискованного поведения; N_i — число эпизодов i-го вида рискованного поведения, в которых принимал участие респондент; $\hat{\Pr}$ — вероятность заражения из-за участия в конкретном виде рискованного поведения; $\Pr$ — общая вероятность заражения респондента. Величины p_i считаются известными [11].

Комбинируя модель Белла–Тревино, уравнение пуассоновского процесса и случайную величину, характеризующую его интенсивность, получим:

$$\hat{\Pr} = 1 - \mathrm{e}^{-\hat{\lambda} p \Delta t},$$

где p_i — вероятность заразиться за один эпизод рискованного поведения; Δt — длительность временнòго интервала, для которого подсчитывается кумулятивный риск заразиться; $\hat{\lambda}$ — интенсивность рискованного поведения. В этом случае точечная оценка вероятности заражения $\Pr = \mathbf{E}\hat{P}_r$. Исследуя свойства случайной величины $\hat{\Pr}$, можно получить и другие оценки вероятности заражения.

Кроме того, мы можем опереться и на грубые оценки L^+, L^-, которые в таком случае равны: $\lambda^+ = \dfrac{\nu}{\tau - \Delta\tau}$ и $\lambda^- = \dfrac{\nu}{\tau + \Delta\tau}$, а грубые верхняя и нижняя оценки вероятности заразится равны: $\Pr_i^+ = 1 - \mathrm{e}^{-\lambda^+ p_i \Delta t}$, $\Pr_i^- = 1 - \mathrm{e}^{-\lambda^- p_i \Delta t}$.

Часть результатов, представленных в настоящей работе, была получена на основе результатов исследований, поддержанных грантом РГНФ «Взаимосвязь адаптивных стилей ВИЧ-инфицированных и степени рискованности их поведения» №07-06-00738а, госконтрактом № 2.442.11.7489, шифр 2006-РИ-19.0/001/209, на НИР «Психологическая защита и копинг-стратегии ВИЧ-инфицированных с точки зрения опасности для общественного здоровья» в рамках ФЦНТП «Исследования и разработки по приоритетным направлениям развития науки и техники на 2002–2006 годы», грантом СПбНЦ РАН на 2007 год «Моделирование и измерение количественных характеристик ВИЧ-рискованного поведения на основе обработки ответов респондентов» № 2-199. Руководитель проектов — Т. В. Тулупьева.

Часть результатов получена в проекте «Оценка вероятности заражения ВИЧ-инфекцией на основе сведений о последних N эпизодах рискованного поведения, а также статистическое моделирование ограниченных указанных серий эпизодов», поддержанном грантом №02/2.1/17-03/48 (в 2007 году) Конкурса для студентов и аспирантов вузов и академических институтов, расположенных на территории Санкт-Петербурга. Руководитель проекта — А. Е. Пащенко.

Литература

1. *Уоткинс П. Дж.* Сахарный диабет. 2-е изд. / Пер. с англ. М.: Изд. БИНОМ, 2006. 134 с.
2. Консультирование по вопросам формирования и поддержания приверженности к антиретровирусной терапии у потребителей инъекционных наркотиков [Электронный ресурс] <http://www.pmplus.org/content/view/139/37/> (По состоянию на 25.12.2009.)
3. Международный альянс по ВИЧ/СПИД в Украине [Электронный ресурс] <http://www.aidsalliance.kiev.ua/> (По состоянию на 25.12.2009.)
4. *Тулупьева Т. В., Тулупьев А. Л., Пащенко А. Е., Красносельских Т. В.* Приверженность ВААРТ и рискованное поведение среди пациентов Санкт-Петербургского Центра СПИД: статистические модели, психологические и социо-демографические факторы //

Труды СПИИРАН. 2008. СПб.: Наука, 2008. Вып. 6. С. 207–237.
5. *Rothman K. J.* Epidemiology: An Introduction. Oxford University Press, 2002. 223 p.
6. *Тулупьева Т. В., Тулупьев А. Л., Столярова Е. В., Пащенко А. Е.* Анализ особенностей рискованного поведения в модели адаптивных стилей ВИЧ-инфицированных (на основе результатов опроса пациентов Санкт-Петербургского СПИД-Центра) // Труды СПИИРАН. 2007. СПб.: Наука, 2007. Вып. 5. С. 117–150.
7. *Тулупьева Т. В., Пащенко А. Е. , Тулупьев А. Л., Красносельских Т. В., Казакова О. С*. Модели ВИЧ-рискованного поведения в контексте психологической защиты и других адаптивных стилей. СПб.: Наука, 2008. 140 с.
8. *Bell D. C., Trevino R. A.* Modeling HIV Risk [Epidemiology] // J. Acquir Immune Defic Syndr. 1999. С. 280–287. vol.22, № 3.
9. *Пащенко А. Е., Тулупьев А. Л., Тулупьева Т. В.* Базисная темпоральная онтология для обработки ответов об участии в рискованном поведении, связанном с передачей ВИЧ // Научная сессия МИФИ-2008. Сб. науч. трудов. В 15 томах. Т. 10. Интеллектуальные системы и технологии. М.: МИФИ, 2008. С. 109–111.
10. *Тулупьева Т. В., Тулупьев А. Л., Пащенко А. Е., Сироткин А. В., Столярова Е. В., Ламанова Е. Б., Бадосова Н. В., Никитин П. В.* Психологическая защита и копинг-стратегии ВИЧ-инфицированных с позиции опасности для общественного здоровья: автоматизация сбора данных и итоги исследования // Труды СПИИРАН. 2007. СПб.: Наука, 2007. Вып. 4. С. 357–387.
11. *Bell D. C., Atkinson J. S., Mosier V., Riley M., Brown V. L.* The HIV Transmission Gradient: Relationship Patterns of Protection // AIDS Behav. 2007. С. 789–811. Vol. 11 № 6.
12. *Хованов Н. В*. Анализ и синтез показателей при информационном дефиците // СПб.: Изд-во СПбГУ, 1996. 196 с.