Математические основы анализа данных. Дизайн исследования

Солодушкин Святослав Игоревич

Кафедра вычислительной математики и компьютерных наук,
УрФУ имени первого Президента России Б.Н. Ельцина

Март 2022

Дизайн исследования

Дизайн исследования — это комбинация требований относительно сбора и анализа данных, необходимых для достижения целей исследования.

Все статистические исследования можно разделить на две группы в зависимости от решаемых задач:

1. описательные,
2. анализирующие причинно-следственные связи
   a. экспериментальные,
   b. наблюдательные (обсервационные).
      i. случай–контроль,
      ii. когортные
      iii. кросс-секционные

Исследование «случай–контроль»

Исследование «случай–контроль» — исследование, при котором изучаемые группы обследуемых объектов набираются по следующему принципу.

В группу I входят объекты, имеющие изучаемое состояние (группа случаев), в группу II — объекты, не имеющие изучаемого состояния (группа контроля).

После набора групп у объектов выясняется наличие изучаемых факторов риска и производится оценка наличия данных факторов в группе случаев и группе контроля.

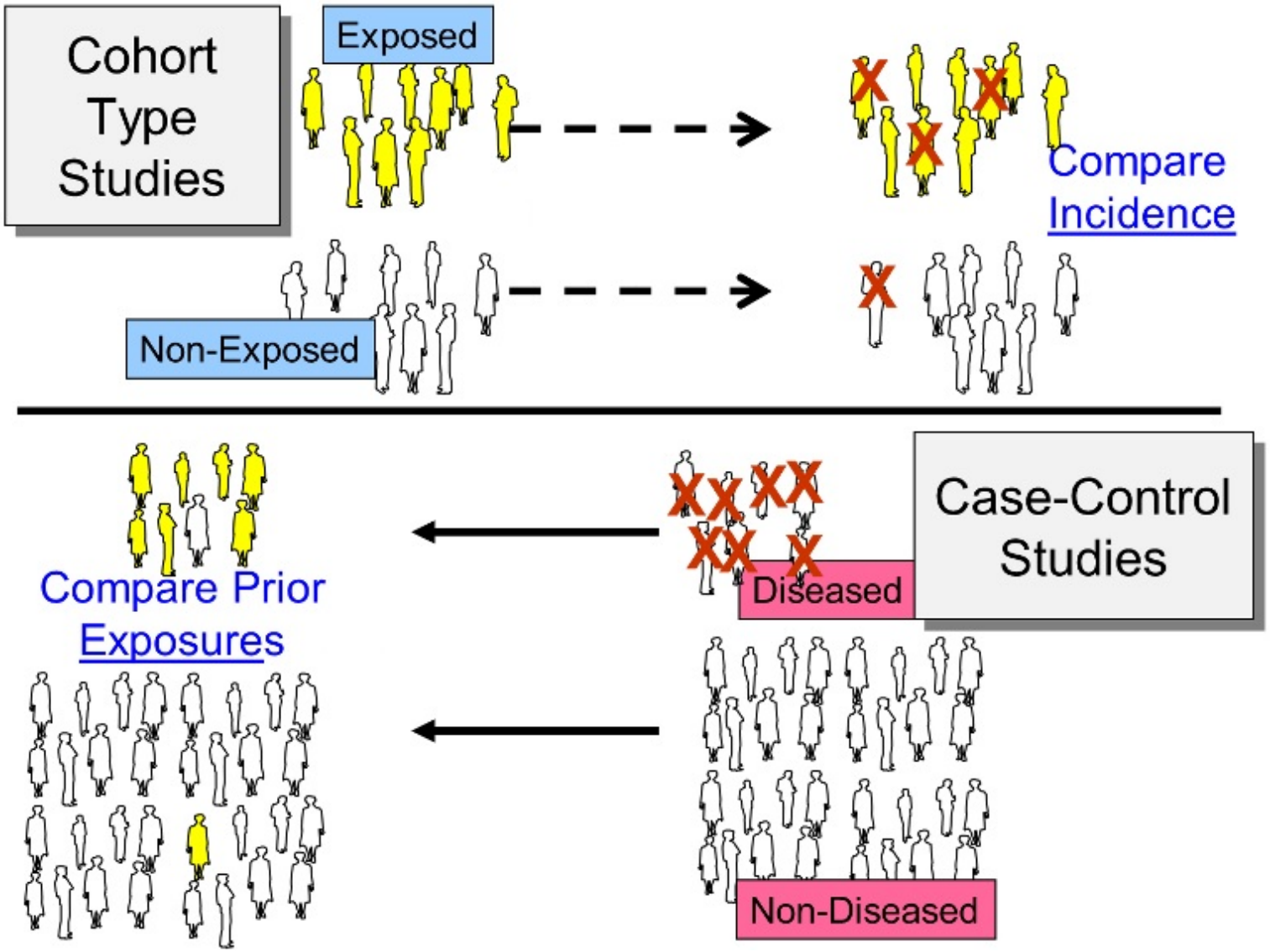
Данное исследование является ретроспективным, так как факт наступления состояния известен, влияние факторов риска зафиксировано уже на этапе начала исследования, а исследователь только собирает эту информацию.

Когортные исследования

Когортные исследования — исследования, при которых изучаемые группы объектов набираются по следующему принципу.

В группу I входят объекты, не имеющие изучаемого состояния, но имеющие изучаемый фактор риска (экспонированная группа), в группу II – объекты, также не имеющие изучаемого состояния, но не имеющие еще и изучаемый фактор риска (неэкспонированная группа).

Дизайны исследований



Пример

Исследуем связь рака легких (среди 40-летних мужчин) и курения (стаж курения с 18 лет, 1 пачка/день).

Пусть курящих в популяции 60%
Среди курящих рак развивается в 90% случаев
Среди некурящих рак развивается в 20% случаев

В популяции, например, из 200 человек (критерии включения см. в первой строке слайда) имеют место следующие пропорции:

	Рак	Нет рака	
Курит	108	12	120
Не курит	16	64	80
	124	76	200

Когортное проспективное исследование

Наберем 200 пациентов (40-летних мужчин, стаж курения с 18 лет, 1 пачка/день). В репрезентативной выборке будут выполнены те же пропорции, что и во всей популяции, следовательно, будем иметь 120 (60%) курящих пациентов, 80 (40%) некурящих.

Сравним количество случаев рака в каждой группе:
90% в группе курильщиков,
40% в группе некурильщиков.

В группе курильщиков рак развивался чаще, значит, он является фактором риска.

Как формально оценить «чаще» или «больше»?

Шанс, отношение шансов

Пусть курящих в популяции 60%. Среди курящих рак развивается в 90% случаев, среди некурящих в — 20%.
В популяции из 200 человек следующие пропорции:

	Рак	Нет рака	
Курит	108	12	120
Не курит	16	64	80
	124	76	200

Шанс рака в группе курильщиков,

$$O(can)_{smoke} = 108/12 = 9.$$

Шанс рака в группе некурильщиков,

$$O(can)_{nonsmoke} = 16/64 = 0.25.$$

Отношение шансов рака (курильщиков VS некурильщиков),

$$OR(can)_{smoke/nonsmoke} = \frac{O(can)_{smoke}}{O(can)_{nonsmoke}} = \frac{108}{12} : \frac{16}{64} = 36.$$

Шанс, отношение шансов

	Рак	Нет рака	
Курит	a	b	a + b
Не курит	c	d	c + d
	a + c	b + d	a+b+c+d

Шанс рака в группе курильщиков,

$$O(can)_{smoke} = a/b.$$

Шанс рака в группе некурильщиков,

$$O(can)_{nonsmoke} = c/d.$$

Отношение шансов рака (курильщиков VS некурильщиков),

$$OR(can)_{smoke/nonsmoke} = \frac{O(can)_{smoke}}{O(can)_{nonsmoke}} = \frac{a}{b} : \frac{c}{d} = \frac{ad}{bc}.$$

Риск, относительный риск

Риск рака в группе курильщиков

$$R(can)_{smoke} = 108/120 = 0.9.$$

Риск рака в группе некурильщиков

$$R(can)_{nonsmoke} = 16/80 = 0.2.$$

Относительный риск рака (курильщиков VS некурильщиков),

$$RR(can)_{smoke/nonsmoke} = \frac{R(can)_{smoke}}{R(can)_{nonsmoke}} = \frac{108}{120} : \frac{16}{80} = 4.5$$

Nota bene, OR(рак) кур/нек = 36, как это эффектнее выглядит, по сравнению с RR = 4.5. А как устрашающе звучит!

Риск: risk и hazard

Риск — это вероятность, он болше 1 быть не может.

В английском языке сущестуют два термина: risk и hazard, оба переводятся как «риск». Прианализе таблиц сопряженности вычисляют risk, он же вероятность. При анализе таблиц выживаемости (англ. survival tables) вычисляют hazard, который может быть больше 1. Ориентироваться надо на контекст.

Достоинства и недостатки когортных исследований

Когортное исследование — единственный способ истинной оценки относительного риска.

1. Число включенных в исследование объектов (респондентов) должно быть значительно больше, чем число объектов с изучаемым исходом (характерно для проспективных когортных исследований).
2. Высокая стоимость исследования из-за того, что приходится исследовать большое число объектов в течение продолжительного времени.
3. Результаты долгое время остаются неизвестными (характерно для проспективных когортных исследований)

Исследование «случай–контроль»

Наберем 100 пациентов с раком и 100 пациентов без рака легких. Сравним количество куращих в каждой группе.

post hoc ergo propter hoc
После этого — значит по причине этого.

В группе курильщиков рак развивался чаще, значит, он является фактором риска.

Мы сами произвольно определили размеры групп, точнее, соотношение размеров групп. Соотношение 50:50 не соответсвует распространенности ракак в популяции.

Отношение шансов в случае–контроле

Пусть курящих в популяции 60%. Среди курящих рак развивается в 90% случаев, среди некурящих в — 20%.
Наберем 100 пациентов с раком и 100 пациентов без рака легких.

	Рак	Нет рака	
Курит	87	16	103
Не курит	13	84	97
	100	100	200

Шанс рака в группе курильщиков (смысла не имеет),

$$O(can)_{smoke} = 87/16 = 5.44.$$

Шанс рака в группе некурильщиков (смысла не имеет),

$$O(can)_{nonsmoke} = 13/84 = 0.15.$$

Отношение шансов рака (курильщиков VS некурильщиков),

$$OR(can)_{smoke/nonsmoke} = \frac{O(can)_{smoke}}{O(can)_{nonsmoke}} = \frac{87}{16} : \frac{16}{84} = 35.$$

Пропорции в случае–контроле

Откуда мы взяли такие пропорции в таблице: 87/13 и 16/84?

Воспользовались формулой Байеса.
Гипотеза $H0$ состоит в том, что пациент не курит, $P(H0) = 0.6$.
Гипотеза $H1$ состоит в том, что пациент курит, $P(H1) = 0.4$.
A — событие, состоящее в том, что у пациента рак.

$$P(A) = 0.9 \cdot 0.6 + 0.2 \cdot 0.4 = 0.54 + 0.08 = 0.62$$

$$P(H0|A) = \frac{P(A|H0)P(H0)}{P(A)} = \frac{0.54}{0.62} \approx 0.87096774.$$

Отношение шансов в «случае–контроле»

Наберем **200** пациентов с раком и 100 пациентов без рака легких.

	Рак	Нет рака	
Курит	174	16	190
Не курит	26	84	110
	200	100	300

Шанс рака в группе курильщиков (смысла не имеет),

$$O(can)_{smoke} = 176/16 = 10.88.$$

Шанс рака в группе некурильщиков (смысла не имеет),

$$O(can)_{nonsmoke} = 26/84 = 0.31.$$

Отношение шансов рака (курильщиков VS некурильщиков),

$$OR(can)_{smoke/nonsmoke} = \frac{O(can)_{smoke}}{O(can)_{nonsmoke}} = \frac{176}{16} : \frac{26}{84} = 35.$$

Отношение шансов в случае–контроле

В исследованиях случай–контроль отношение шансов можно считать. Оно является несмещенной оценкой отношения шансов в генеральной совокупности.

Относительный риск в случай-контроле

Наберем 100 пациентов с раком, 100 пациентов без рака.

	Рак	Нет рака	
Курит	87	16	103
Не курит	13	84	97
	100	100	200

Риск рака в группе курильщиков

$$R(can)_{smoke} = 87/103 = 0.84.$$

Риск рака в группе некурильщиков

$$R(can)_{nonsmoke} = 13/97 = 0.13.$$

Относительный риск рака (курильщиков VS некурильщиков),

$$RR(can)_{smoke/nonsmoke} = \frac{R(can)_{smoke}}{R(can)_{nonsmoke}} = \frac{87}{103} : \frac{13}{97} = 6.3$$

Не совпадает с $RR = 4.5$, найденным в когортном исследовании.

Относительный риск в случай-контроле

Наберем 200 пациентов с раком, 100 пациентов без рака.

	Рак	Нет рака	
Курит	174	16	190
Не курит	26	84	110
	200	100	300

Риск рака в группе курильщиков

$$R(can)_{smoke} = 174/190 = 0.91.$$

Риск рака в группе некурильщиков

$$R(can)_{nonsmoke} = 26/110 = 0.24.$$

Относительный риск рака (курильщиков VS некурильщиков),

$$RR(can)_{smoke/nonsmoke} = \frac{R(can)_{smoke}}{R(can)_{nonsmoke}} = \frac{174}{190} : \frac{26}{110} = 3.87$$

И при увеличении объема группы онкологических пациентов этот «относительный риск» стремится к единице. . .

Относительный риск в случай-контроле

В исследованиях «случай–контроль» относительный риск считать нельзя. Он лишен содержательного смысла.

Преимущества исследования «случай–контроль»

1. Экономичность.
2. Быстрота получения результатов.
3. Возможность изучения редких событий.
4. Возможность изучать большой спектр факторов риска.
5. В случае адекватного подбора контрольной группы мало отличаются по своей ценности от когортных исследований.
6. Отсутствует потеря наблюдаемых лиц в ходе исследования.

Недостатки исследований «случай–контроль»

1. Сложность подбора контрольной группы.
2. Не подходят для изучения редких факторов риска.
3. Имеют ограниченные возможности установления временной последовательности событий.
4. Трудность получения достоверной информации об уровне воздействия на индивида с течением времени.

Кросс-секционные исследования

Кросс-секционные (поперечные) исследования проводятся одномоментно для выяснения распространенности факторов риска и исходов. Необходимо отметить, что данное исследование, как правило, не проводится для выяснения причинно-следственной связи между факторами риска, лечением, исходами и т. д.

Задания

1. Сформулируйте пример задачи, которую можно было бы решать в рамках Кросс-секционного дизайна исследования.
2. Можно ли оценивать относительный риск в кросс-секционных исследованиях? Ответ обоснуйте.
3. Проводится клиническое исследование результатов эндоваскулярной коррекции стеноза почечной артерии. Сформировано две группы: в группе исследования пациентам проводится ангиопластика, в группе контроля пациенты получают только медикаментозную терапию. В группу исследования попали пациенты со стенозом более 70%, а группу контроля — оставшиеся пациенты. Можно ли назвать данное исследование рандомизированным?