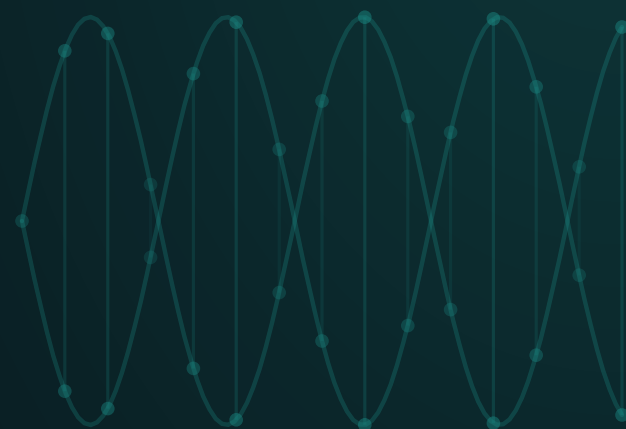


레슨: 유전 연구 결과가 다른 집단에 적용되지 않을 때



유전학

확률과 통계

사회

보건

만 16세+

중급

이 사이트 자체를 데이터로 삼아, 데이터를 비판적으로 읽는 법을 다루는 무료 Lesson입니다. 학급이 같은 예측을 다섯 집단에서 계산하고 얼마나 빗나가는지 쟁니다 — 어떤 변이는 6점, 어떤 변이는 52점입니다. 그다음 저희 자신의 감사 결과를 공개합니다. 여기 실린 변이의 72.2%가 전부 유럽계인 표본에서 발견되었습니다.

자유롭게 읽고, 쓰고, 고쳐 쓰실 수 있습니다. 계정도, 구매도, 수집되는 정보도 없습니다.

전체 레슨과 그림, 갱신본: <https://www.mygenelog.com/ko/updates/lesson-does-this-number-travel>

이 자료는. 데이터를 비판적으로 읽는 법을 다루는 Lesson입니다. 실제 유전학 데이터와, 이 사이트가 스스로를 감사한 결과로 만들었습니다. 어느 학교·칼리지·대학에서나 자유롭게 쓰고 고치실 수 있습니다. 저희에게서 살 것도, 계정도, 누구에게서든 수집하는 데이터도 없습니다.

- 수준: 만 16세부터 학부까지. 유전자, 대립유전자, 유전형이라는 말과 소수의 제곱을 계산할 수 있으면 됩니다.
- 시간: 50-60분. 계산은 15분쯤 걸리고, 끝의 논쟁은 남은 시간을 다 쓸 수 있으며 대개 그래도 좋습니다.
- 준비물: 계산기나 휴대전화, 적을 것. 숫자는 아래에 전부 적혀 있으므로 교실에 인터넷이 없어도 됩니다.
- 다루는 내용: 대립유전자 빈도, 보유 비율과 산식이 생물학에 따라 달라지는 이유, 집단 구조, 참고치란 무엇이고 어떻게 어긋나는지, 실제 연구 분야의 표본 편향, 그리고 데이터를 믿는 대신 감사하는 법.

여기서 시작합니다: 정상인 결과를 질병이라고 불렀다

거의 모두가 받아 본 혈액검사가 있습니다. 세균 감염에 대응하는 백혈구인 호중구의 수를 세는 것입니다. 그 수치가 참고치 아래로 나오면 그것이 호중구감소증이고, 가볍게 넘기지 않습니다. 약을 끊어야 한다는 뜻일 수도, 의료 현장의 업무를 멈춰야 한다는 뜻일 수도, 항암제 용량을 줄여야 한다는 뜻일 수도, 골수 검사를 해야 한다는 뜻일 수도 있습니다.

그런데 문제가 있습니다. 흑인 미국인의 약 3분의 2, 그리고 서·중·동아프리카 출신 사람의 대다수가 아무도 아프게 하지 않으면서 혈액의 호중구 수를 낮추는 변이를 가지고 있습니다. Duffy-null이라고 하며, 세계에서 가장 흔한 변이 속에 들고, 한 종류의 말라리아로부터 보호해 주기 때문에 흔해졌습니다.

수십 년 동안 그 수치를 견주던 참고치는 대체로 이 변이를 가지지 않은 사람들로 만들어져 있었습니다. 건강한 사람들이 백혈구 수가 낮다는 말을 들었습니다. 그것 때문에 치료가 바뀌거나 보류된 사람도

있었습니다.

추론을 틀린 사람은 없습니다. 참고치는 의도된 대로 정확하게 적용되었습니다. 다만 그것이 옮겨 가지 않았을 뿐입니다 — 그리고 앞으로 한 시간은, 어떤 숫자를 사람에게 쓰기 전에 그것이 옮겨 가는지 확인하는 방법에 대한 것입니다.

이렇게 하세요

아래 숫자는 전부 공개된 대립유전자 빈도에서 온 실제 값입니다. 교실 안의 누구도 검사받거나, 세어지거나, 자기 자신에 대해 질문받지 않습니다.

1. 먼저 예측을 받으십시오. “어떤 나라의 연구가 25%의 사람이 어떤 변이를 가지고 있다고 밝혔습니다. 다른 나라를 가서 재면 그 25%가 얼마나 빛나갈 수 있을까요? 몇 점? 10점? 그 이상?” 답을 칠판에 적으십시오.
2. 도구를 알려 주십시오. 어떤 대립유전자의 빈도가 q 일 때, 한 개 이상 가진 사람의 비율은 $1 - (1 - q)^2$ 입니다. 하나를 같이 해 보십시오. $q = 0.131$ 이면 $1 - 0.869^2 = 1 - 0.755 = 24.5\%$ 입니다.
3. 대조군부터 시작하고, 그것이 대조군이라고 말하십시오. rs2075650의 빈도를 나눠 주십시오. 유럽계 0.131, 동아시아계 0.097, 아프리카계 0.131, 남아시아계 0.125, 아메리카계 0.105. 각 모음이 집단별 보유 비율을 계산합니다. 답은 18.5%에서 24.5% 사이에 들어오고, 최악의 경우가 6점입니다. 유럽에서 만든 예측이 어디에서나 대체로 맞습니다. 이 단계는 결가지가 아닙니다. 이것이 없으면 학급은 연구 결과가 결코 옮겨 가지 않는다고 결론짓게 되는데, 그것은 틀렸고 들어올 때 가지고 있던 생각보다 나쁜 결론입니다.
4. 이제 두 번째 변이입니다. VKORC1의 rs9923231 — 와파린이 차단하는 효소의 유전자입니다. 빈도: 유럽계 0.388, 동아시아계 0.885, 아프리카계 0.054, 남아시아계 0.145, 아메리카계 0.411. 같은 계산입니다.
5. 그다음 오차를 재십시오. 이것이 실제 과제입니다. 유럽계의 답을 가져다 다른 집단들의 예측값으로 씁니다. 빼십시오. 매번 몇 퍼센트포인트나 틀립니까? 가장 큰 것을 1단계의 추측 옆에 적으십시오.
6. 이야기에 나온 것으로 마무리합니다. rs2814778, Duffy-null: 유럽계 0.006, 동아시아계 0.000, 아프리카계 0.964, 남아시아계 0.000, 아메리카계 0.078. 여기서 유럽에서 만든 기대치가 아프리카계 집단에 대해 무엇이라고 말할지 물으십시오.
7. 마지막 변이에는 한 단계가 더 있고, 좋은 단계입니다. Duffy-null은 두 개를 가진 사람에게서만 혈구 수를 바꿉니다. 그러므로 맞는 계산은 $1 - (1 - q)^2$ 가 아니라 q^2 입니다. 계산해 보십시오. 아프리카계 표본에서 $0.964^2 = 92.9\%$, 유럽계 표본에서 $0.006^2 = 0.004\%$ 입니다. 왜 식이 바뀌었는지 물으십시오. 답은 수학이 아니라 생물학이 바뀌었기 때문이고, 그것을 말할 수 있는 학생은 이 수업에서 가장 널리 쓰이는 생각을 가져간 것입니다.

교실을 데이터로 만들지 마십시오. 이 수업은 집단에 대한 것이고, 뻔한 실패는 학생들을 배경으로 나누어 세는 것입니다. 학생에게 가족이 어디 출신인지 묻지 말고, 그것으로 묵지 말고, 학급을 어떤 집단의 예시로 쓰지 마십시오. 교실은 표본이 아니고, 학생의 조상은 교보재가 아닙니다.

이 사이트의 모든 수업과 같습니다. 어떤 경로로든, 누구에게서든 유전·건강 데이터를 수집하지 마십시오. 이 수업의 숫자는 공개된 집단 빈도이고, 그것이면 충분합니다.

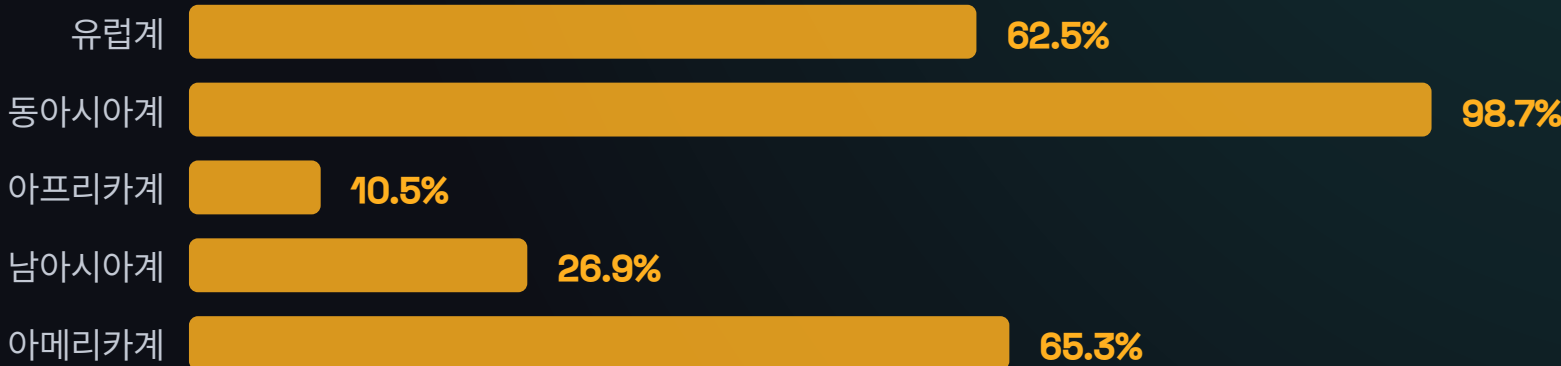
학생들이 방금 찾아낸 것

한 개 이상 가진 사람의 비율 — 공개된 대립유전자 빈도에서 계산

rs2075650 — 대조군: 이것은 옮겨 갑니다



rs9923231 — VKORC1, 와파린의 표적: 이것은 옮겨 가지 않습니다



rs2814778 — Duffy-null: 이것은 거의 한 곳에만 있습니다



Ensembl을 통한 1000 Genomes phase 3 — 보유 비율은 $1 - (1 - q)^2$ 로, 인용이 아니라 여기서 계산했습니다

변이 셋, 계산 하나, 그리고 “한 곳에서 나온 숫자를 다른 곳에 써도 되는가”에 대한 완전히 다른 세 개의 답. 맨 위가 이것을 고발이 아니라 질문으로 만들어 줍니다. 잘 옮겨 가는 경우도 있습니다.

가운데 변이를 두고 잠깐 머물 값어치가 있습니다. 유럽에서 만든 예측은 동아시아계 집단에 대해 36점 낮고, 아프리카계 집단에 대해 52점 높습니다 — 똑같이 올바른 계산을, 다른 사람들에게 적용한 결과로 말입니다.

그리고 이것은 추상적인 변이가 아닙니다. 와파린의 표적 유전자에 있고, 와파린은 너무 적은 것과 너무 많은 것 사이의 간격이 좁아서 용량을 찾는 동안 반복해서 검사를 받아야 하는 약입니다.

실제 데이터: 이 수업을 위해 저희를 감사했습니다

이것을 남이 저지른 잘못으로 가르치기는 쉽습니다. 그래서 같은 질문을 저희 목록에 던졌습니다.

이 사이트의 모든 변이는 출처 논문을 밝히고 있습니다. 그 논문 187편 전부를 가져와, 각 연구의 발견 표본에 실제로 어떤 조상 집단이 들어 있었는지 GWAS Catalog에 물어 세었습니다.

이 감사는 2026년 9월 8일에, 그때 이곳에 발행돼 있던 변이 608건을 대상으로 했습니다. 목록은 그 뒤로 자랐고 이 수치들은 일부러 함께 갱신하지 않습니다. 아래의 백분율은 그 논문 집합을 기술한 것이고, 더 큰 숫자로 바꿔 넣으면 그 옆의 모든 비율이 한 번도 센 적 없는 무언가를 기술하게 됩니다.

이 사이트의 변이 608건, 발견된 표본의 조상 구성별

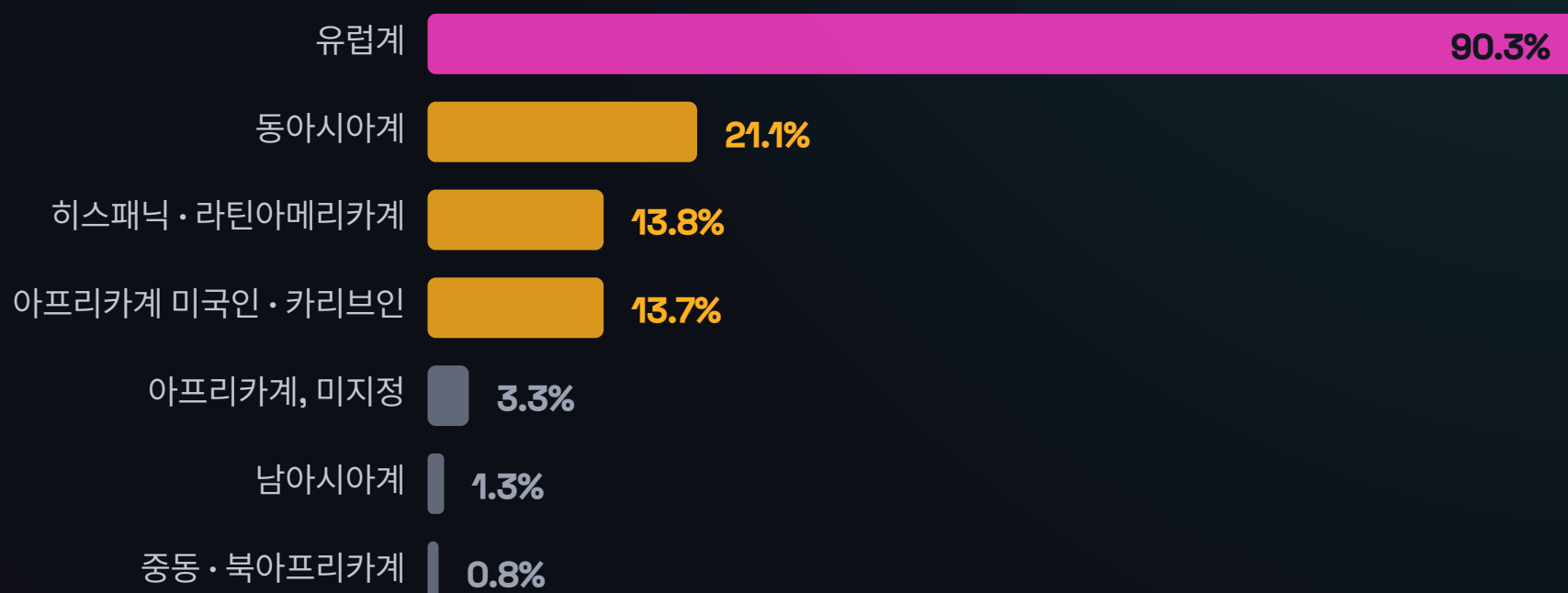


- **439**건 — 전부 유럽계인 표본에서 발견
- **110**건 — 유럽계를 포함한 혼합 표본에서
- **59**건 — 유럽계가 전혀 없는 표본에서, 9.7%

저희 목록을, 그 뒤에 있는 논문 187편 전부에 대해 GWAS Catalog의 조상 기록과 대조해 직접 감사했습니다. 내놓고 싶은 숫자는 아니고, 그것이 내놓는 이유의 대부분입니다.

이 사이트에 공개된 변이의 **72.2%**가 전부 유럽계인 표본에서 발견되었습니다. **18.1%**가 유럽계를 포함한 혼합 표본에서 나왔습니다. 유럽계 조상이 전혀 없는 연구에서 나온 것은 열에 하나가 되지 않습니다.

각 집단이 발견 표본에 포함된 변이의 비율 (한 연구가 여러 집단을 포함할 수 있습니다)



남아시아는 지금 살아 있는 사람의 약 4분의 1입니다. 이 그래프에서는 1.3%입니다.

같은 감사를 다른 방향으로 자른 것입니다. 한 연구가 여러 집단을 모집할 수 있어서 합이 100%를 넘습니다 — 앞 그림의 두 번째 막대가 그것이었습다.

다른 방향으로 자르면 더 뚜렷합니다. 남아시아계 조상은 저희 변이 608건 중 8건의 발견 표본에 들어 있습니다 — **1.3%**입니다. 남아시아는 지금 살아 있는 사람의 약 4분의 1입니다.

이것은 저희가 만든 문제가 아니고 저희 힘만으로 고칠 수도 없습니다. 저희는 연구된 것을 싣고, 연구된 것이 이렇습니다. 다만 남의 사각지대를 설명하면서 자기 것을 감추는 사이트는 아무것도 가르치고 있지 않으므로, 이 숫자를 페이지에 올립니다. 학급이 확인하고 싶다면 모든 변이 페이지에 출처가 적혀 있고 GWAS Catalog는 공개되어 있습니다.

아직 끝나지 않은 논쟁

불균형이 문제라는 데는 모두가 동의합니다. 무엇을 할 것인가는 정말로 갈리고, 학급이 그 논쟁 전체를 다룰 수 있습니다.

“더 넓게 모집하라.” 뻔한 답이고, 길게 보면 옳은 답입니다. 동시에 역사가 있는 답이기도 합니다. 연구가

지역사회를 상대로, 그들의 이해나 이익 없이 이루어진 일이 충분히 많았기 때문에 그 의심은 비합리적인 것이 아니라 얻어진 것입니다. 진지한 답이라면 몇 명을 등록시키는지만이 아니라 참여자에게 무엇을 알리고, 그들이 무엇을 얻고, 그 뒤에 시료를 누가 통제하는지를 포함해야 합니다.

“대신 더 나은 방법을 쓰라.” 통계적 방법으로 집단 사이에서 정보를 빌려 와, 새로 모집하지 않고도 격차를 일부 메울 수 있다는 주장이 있습니다. 다른 쪽은 애초에 수집되지 않은 정보를 방법이 만들어 낼 수는 없으며, 이 주장은 대체로 비싼 일을 하지 않기 위한 방편이라고 답합니다.

“그러면 적용되지 않는 곳에는 쓰지 말라.” 논리적이고, 가시가 있습니다. 엄격하게 적용하면 그것은 지금까지 가장 덜 연구된 집단에게 유전적 위험 평가 도구를 주지 않는다는 뜻이 되고, 첫 번째 불이익 위에 두 번째 불이익을 얹는 일입니다.

학급에 건네줄 합의는 없습니다. 세 입장을 주고, 각각에 대가가 있다는 것을 직접 찾아내게 하십시오.

질문

몸풀기

1. rs2075650에는 흥미로운 일이 아무것도 일어나지 않는데, 왜 이 수업에 들어 있습니까?
2. 혈액검사의 참고치는 건강한 사람들의 측정값으로 만듭니다. 완벽하게 올바르게 만든 참고치가 특정 환자에게 틀린 답을 줄 수 있는 경우를 하나 말하십시오.
3. 계산한 것 중 가장 큰 전이 오차는 몇 퍼센트포인트였습니까? 칠판의 추측과 비교하면 어떻습니까?

본문제

1. Duffy-null 계산에는 왜 q^2 를 쓰고 나머지 둘에는 $1 - (1 - q)^2$ 를 썼는지 설명하십시오. 새 변이에서 둘 중 하나를 고르려면 무엇을 알아야 합니까?
2. Duffy-null은 한 종류의 말라리아로부터 보호해 주기 때문에 흔해졌습니다. 어떤 변이가 크게 이롭기도 하면서 세계의 한 부분에만 몰려 있을 수 있는 이유를 자기 말로 설명하십시오.
3. 이 사이트 변이의 72.2%가 전부 유럽계인 표본에서 발견되었습니다. 그 결과를 두 가지 말하십시오. 하나는 예측의 정확도에 대한 것으로, 하나는 정확도와 전혀 상관없는 것으로.
4. 어떤 연구가 참여자의 30%에게 그 변이가 있다고 보고합니다. 그 숫자를 다른 곳에 쓰기 전에 가장 먼저 알아야 할 것 하나는 무엇입니까?

더 나아가기

1. 이 그림들의 이름표 — 유럽계, 동아시아계, 아프리카계 — 는 아주 많은 변이를 담고 있는 그릇입니다. 이 사이트의 앞선 수업에서는 동아시아 집단 둘이 한 변이에서 다섯 배 차이가 났습니다. 그러면 이 이름표는 쓸모없는 것입니까, 쓸모는 있지만 부족한 것입니까? 답을 변호하고, 대신 무엇을 쓰겠는지 말하십시오.
2. 어떤 참고치가 특정 집단을 위한 별도 판을 필요로 하는지 판단할 때 병원이 실제로 따를 수 있는 규칙을 설계하십시오. 화요일 아침에 적용할 수 있을 만큼 구체적이어야 하고, 모든 환자에게 조상을 묻지 않아야 합니다.
3. 위의 세 입장을 가져가십시오. 각각에 대해, 그것을 채택하면 누가 대가를 치르고 채택하지 않으면 누가 대가를 치르는지 말하십시오. 셋 중 무엇을 변호하겠으며, 거기에 대한 가장 강한 반론은 무엇입니까?
4. 이 사이트는 스스로를 감사해 자기에게 유리하지 않은 숫자를 공개했습니다. 그것이 옳은 선택이었습니까? 옳지 않았다는 가장 좋은 주장을 만들어 보고 — 그다음 거기에 답하십시오.

교사용 메모 — 각 질문이 향하는 곳

1. 대조군입니다. 잘 옮겨 가는 사례가 없으면 수업이 “유전학은 결코 일반화되지 않는다”로 지나가 버리는데, 그것은 틀렸고 도움도 되지 않습니다. 방법은 시시한 답도 낼 수 있어야 합니다.
2. 참고치를 만든 건강한 사람들이, 측정에 영향을 주는 무언가에서 그 환자와 체계적으로 다를 때입니다. Duffy-null이 가장 깨끗한 예이고, 나이·고도·운동 훈련을 드는 학생도 있을 텐데 전부 좋은 답입니다.
3. 최악은 rs9923231에서 아프리카계 대 유럽계의 52점입니다. 대부분의 학급은 10점 이하로 추측합니다. 두 숫자를 나란히 적어 두고 지우지 마십시오.
4. 한 개로 충분할 때는 보유 비율이 맞고, 두 개가 필요할 때는 q^2 가 맞습니다. 고르려면 생물학 — 그 효과가 우성인지 열성인지 — 을 알아야 하고, 대립유전자 빈도는 아무리 많아도 그것을 말해 주지 않습니다. 이 수업 전체에서 가장 널리 쓰이는 요점입니다.
5. 선택압이 있는 곳에서 선택이 작동합니다. 어떤 기생충으로부터 보호해 주는 변이는 그 기생충이 있는 곳에서 흔해지고 없는 곳에서는 드물게 남습니다. 잘하는 학생은 이것을 유당분해효소 지속성과 연결하고, 빈도 지도가 왜 전혀 다른 무언가의 지도처럼 보이는지도 알아챌 것입니다.
6. 정확도: 예측과 위험 점수가 발견된 집단 밖에서 덜 믿을 만해집니다. 정확도가 아닌 것: 누구의 질문이 던져지는가, 어떤 질환이 연구되는가, 그 결과로 나온 약에서 누가 이익을 얻는가, 그리고 자기를 연구하지 않은 분야를 누가 믿어 줄 것으로 기대되는가. 잘 논증된 두 번째 결과라면 무엇이든 받아 주십시오.
7. 참여자가 누구였는가. 표본 크기, p-값, 효과 크기를 포함한 나머지 전부는 그 숫자가 옮겨 갈 수 있는가보다 뒤입니다.
8. 쓸모는 있지만 부족합니다. 정직한 답은 이름표가 진짜 중요한 것의 대리물이라는 것이고, 진짜 중요한 것은 보고된 것이 아니라 측정된 유전적 조상입니다. 잘하는 학생은 측정된 조상을 직접 쓰자고 제안할 텐데 — 요즘 연구가 실제로 하는 일입니다 — 그러면 모두의 유전형을 분석해야 한다는 문제에 부딪히게 됩니다.
9. 여기서 가장 어려운 실무 질문입니다. 환자를 인종으로 분류하도록 슬며시 요구하는 규칙을 잘 보고 밀어붙이십시오. Duffy-null의 경우는 이제 배경을 묻는 대신 그 변이 자체를 검사해 처리하는 일이 많은데, 훨씬 나은 규칙이고 학급이 스스로 도달할 수 있는 규칙입니다.
10. 정답이 없습니다. 세 경우 모두 대가가 고르지 않게 분포하고, 그것을 알아채는 것이 결과입니다.
11. 반대편의 가장 강한 주장: 약점을 공개하면 그 사실이 맥락 없이 인용될 빌미가 되고, 여전히 최선인 데이터에 대한 신뢰를 떨어뜨릴 수 있다는 것. 저희가 내놓을 답은, 그 숫자가 참이고 같은 공개 API를 쓰는 누구나 찾아낼 수 있으며, 감춘 뒤에 독자가 스스로 발견한다면 저희에 대해 훨씬 나쁜 것을 알게 되리라는 것입니다. 학급이 저울질하게 두십시오.

더 나아가기

- Duffy-null — 말라리아와 얽힌 역사, 혈구 수치, 그리고 이것이 왜 질환이 아닌지에 대한 전체 페이지.
- 와파린 용량 편차 — VKORC1이 임상에서 왜 중요한지, 그리고 유전형이 무엇을 바꾸고 무엇을 바꾸지 않는지.
- Lesson: 글자 하나, 소단위 넷 — 짝 수업입니다. 교과서라면 같은 칸에 넣었을 두 집단이 다섯 배 차이로 갈라집니다.
- 데이터 페이지 — 실시간 집계, 그리고 실제로 이 사이트를 읽고 있는 것이 무엇인지.

자유롭게 쓰십시오. 인쇄하고, 복사하고, 잘라 쓰고, 학습지에 옮기고, 번역하고, 질문을 바꾸십시오. 허락받을 것도 낼 돈도 없습니다. 출처를 밝히신다면 mygenelog.com이면 충분합니다.

수업에 써 보시고 잘 안 되는 부분이 있으면 알려 주십시오 — 저희에게는 감사 인사보다 그 편이 값집니다.