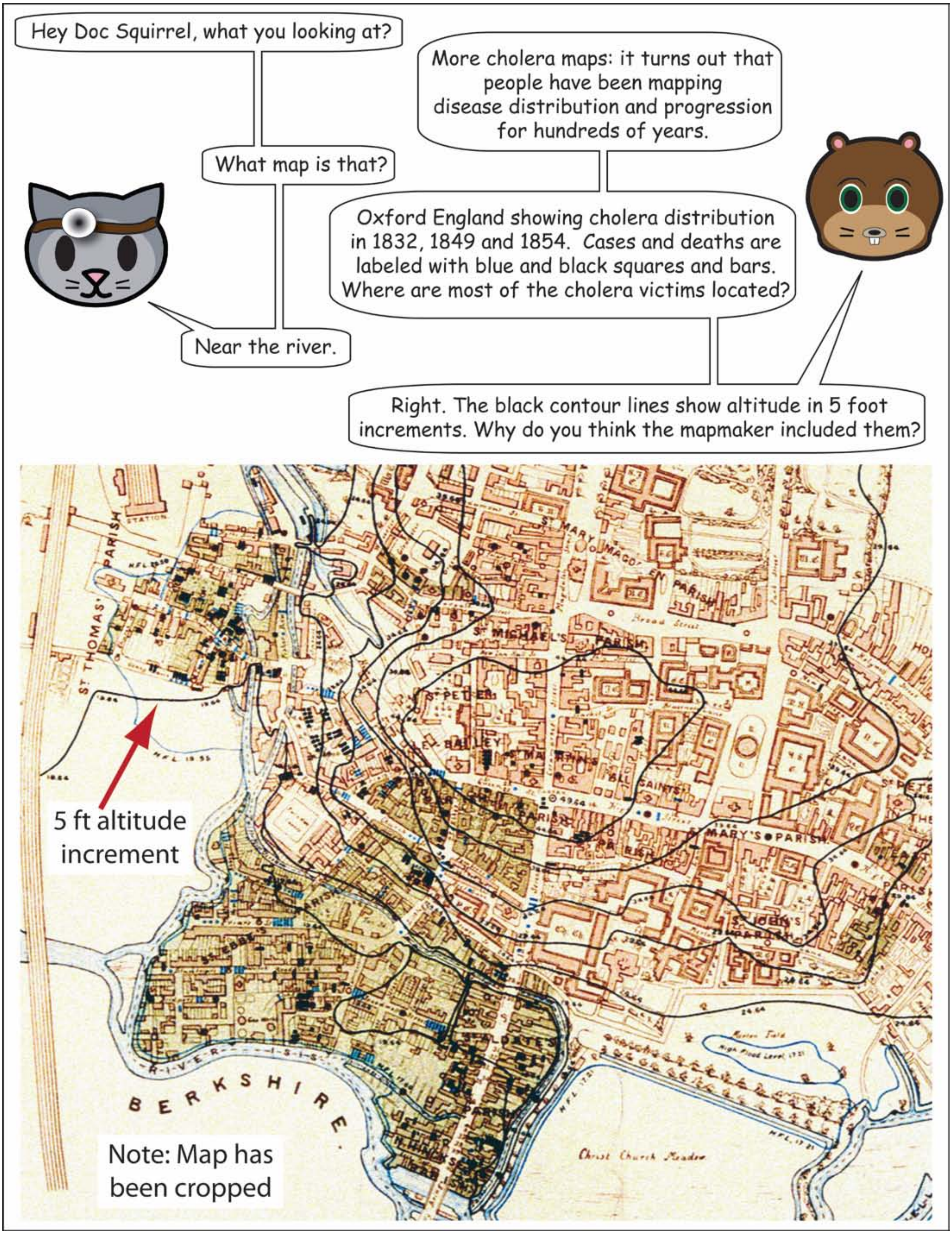
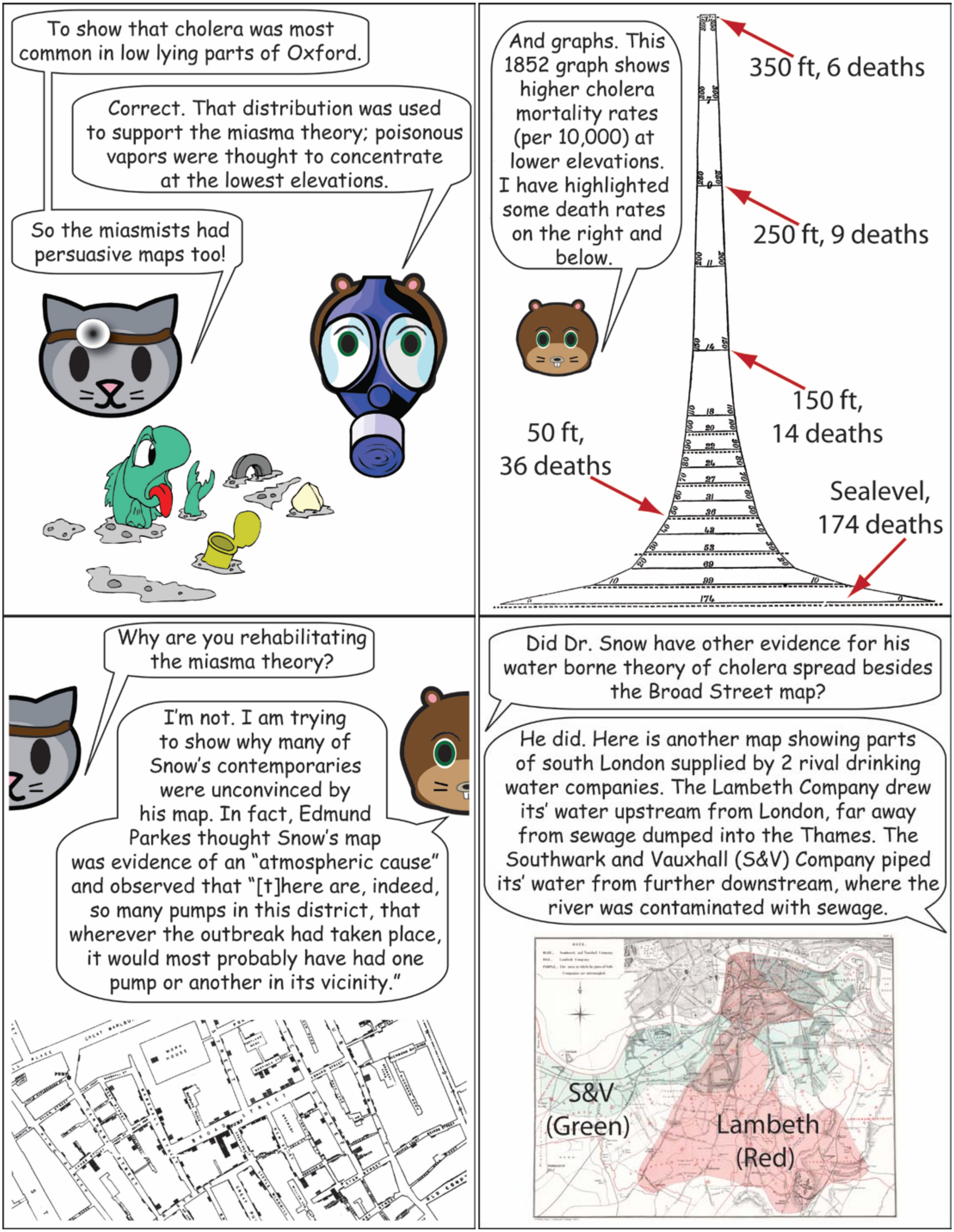
Another “Doc Squirrel” and “Kid Cat” Adventure!!

Cohorts

By Stefan Tigges MD MSCR and David Schulman MD MPH



Hey Doc Squirrel, what you looking at?
More cholera maps: it turns out that people have been mapping disease distribution and progression for hundreds of years.
What map is that?
Oxford England showing cholera distribution in 1832, 1849 and 1854. Cases and deaths are labeled with blue and black squares and bars. Where are most of the cholera victims located?
Near the river.
Right. The black contour lines show altitude in 5 foot increments. Why do you think the mapmaker included them?
STATION
ST. THOMAS' PARISH
ST. MARY MAGDALEN PARISH
ST. MICHAEL'S PARISH
Broad Street
ST. PETER
ST. MARTIN'S
ALL SAINTS
ST. MARY'S PARISH
ST. JOHN'S PARISH
ST. EBBE
ST. ALDATES
RIVER ISIS
BERKSHIRE
Christ Church Meadow
High Flood Level 1821
5 ft altitude increment
Note: Map has been cropped


To show that cholera was most common in low lying parts of Oxford.
Correct. That distribution was used to support the miasma theory; poisonous vapors were thought to concentrate at the lowest elevations.
So the miasmists had persuasive maps too!
And graphs. This 1852 graph shows higher cholera mortality rates (per 10,000) at lower elevations. I have highlighted some death rates on the right and below.
350 ft, 6 deaths
250 ft, 9 deaths
150 ft, 14 deaths
50 ft, 36 deaths
Sealevel, 174 deaths
Why are you rehabilitating the miasma theory?
I'm not. I am trying to show why many of Snow's contemporaries were unconvinced by his map. In fact, Edmund Parkes thought Snow's map was evidence of an "atmospheric cause" and observed that "[t]here are, indeed, so many pumps in this district, that wherever the outbreak had taken place, it would most probably have had one pump or another in its vicinity."
Did Dr. Snow have other evidence for his water borne theory of cholera spread besides the Broad Street map?
He did. Here is another map showing parts of south London supplied by 2 rival drinking water companies. The Lambeth Company drew its' water upstream from London, far away from sewage dumped into the Thames. The Southwark and Vauxhall (S&V) Company piped its' water from further downstream, where the river was contaminated with sewage.
S&V (Green)
Lambeth (Red)

	Lambeth	S&V
Cholera Deaths	461	4,093
Population	173,748	266,516
Mortality Rate per 1000	2.7	15.4

This table compares cholera deaths, total population, and cholera mortality rates of Lambeth and S&V water drinkers for a 14 week period in 1854. What do you think?

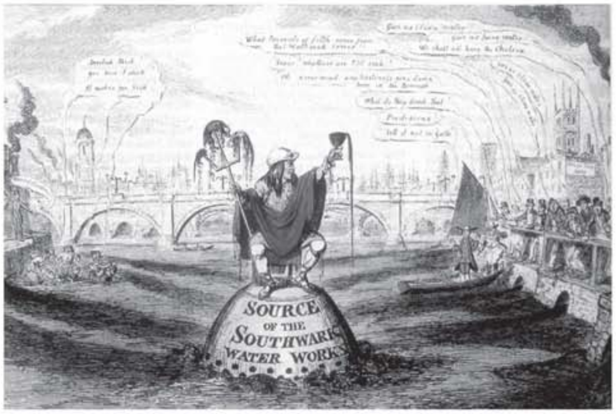
S&V customers are dying of cholera at much higher rates than Lambeth customers.

Correct. This type of study is called a cohort study. We compare outcomes in 2 (or more) groups or populations at risk of developing a disease; an exposed group and an unexposed group.

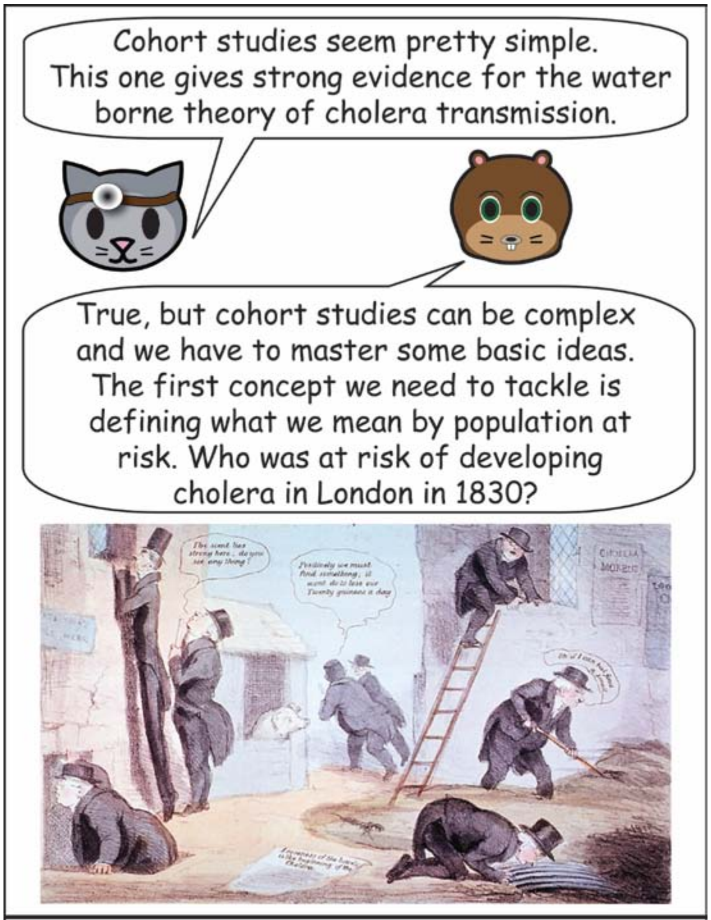
Exposed healthy people → Time → Number of diseased people

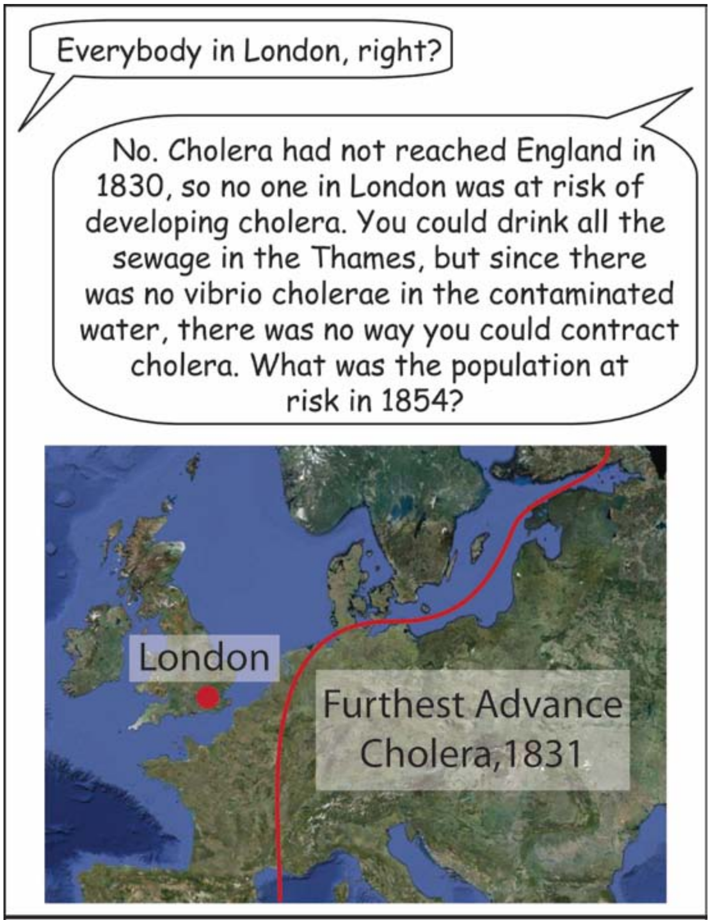
Unexposed healthy people → Time → Number of diseased people

Exposed to what?

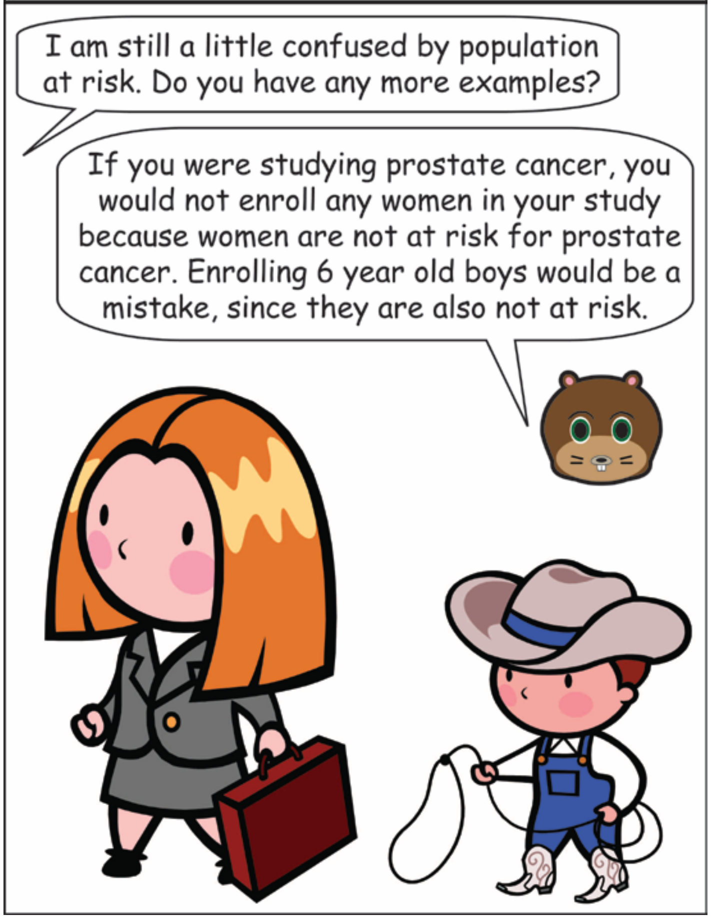


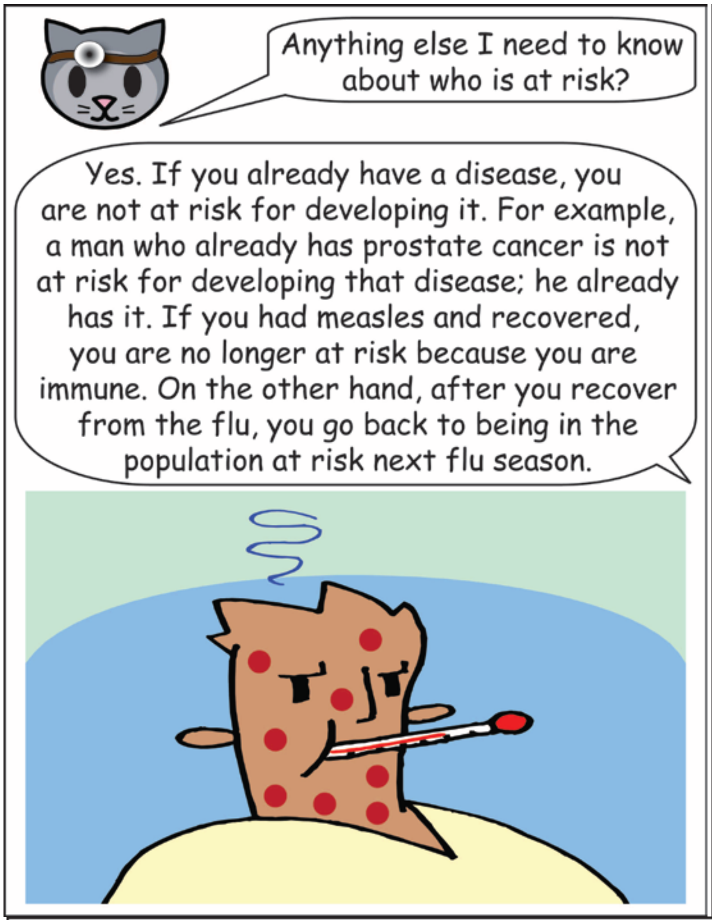
In this case, S&V customers were exposed to water contaminated by sewage. Since Lambeth customers drank uncontaminated water, they are considered unexposed. The exposed group (S&V customers) had an almost 6-fold higher cholera mortality rate than the unexposed (Lambeth customers) group.


Cohort studies seem pretty simple. This one gives strong evidence for the water borne theory of cholera transmission.
True, but cohort studies can be complex and we have to master some basic ideas. The first concept we need to tackle is defining what we mean by population at risk. Who was at risk of developing cholera in London in 1830?
The scent has strong here.. do you see any thing!
Positively we must find something; it wont do to lose our Twenty guineas a day
CHOLERA MORBUS

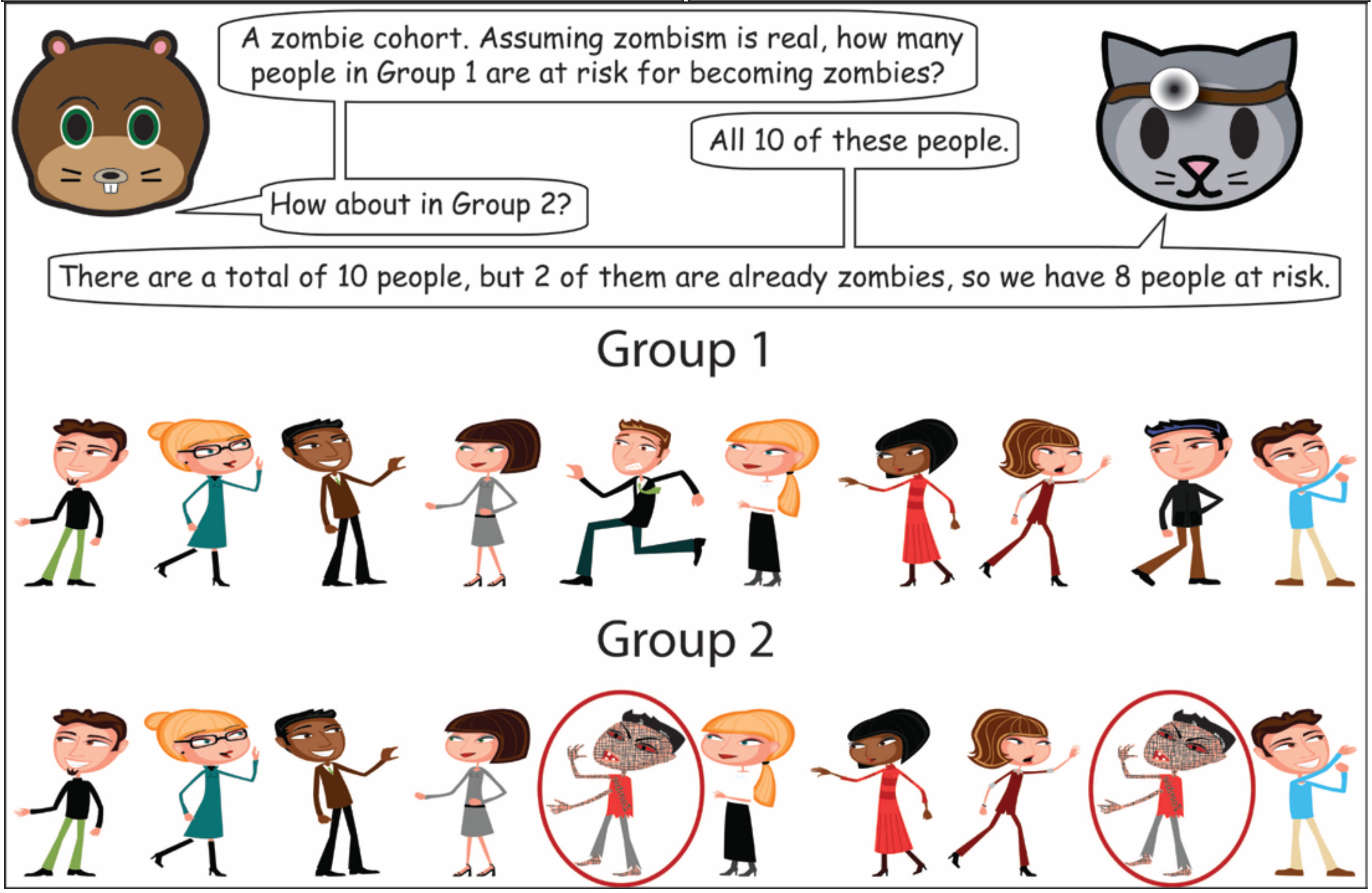

Everybody in London, right?
No. Cholera had not reached England in 1830, so no one in London was at risk of developing cholera. You could drink all the sewage in the Thames, but since there was no vibrio cholerae in the contaminated water, there was no way you could contract cholera. What was the population at risk in 1854?
London
Furthest Advance Cholera, 1831

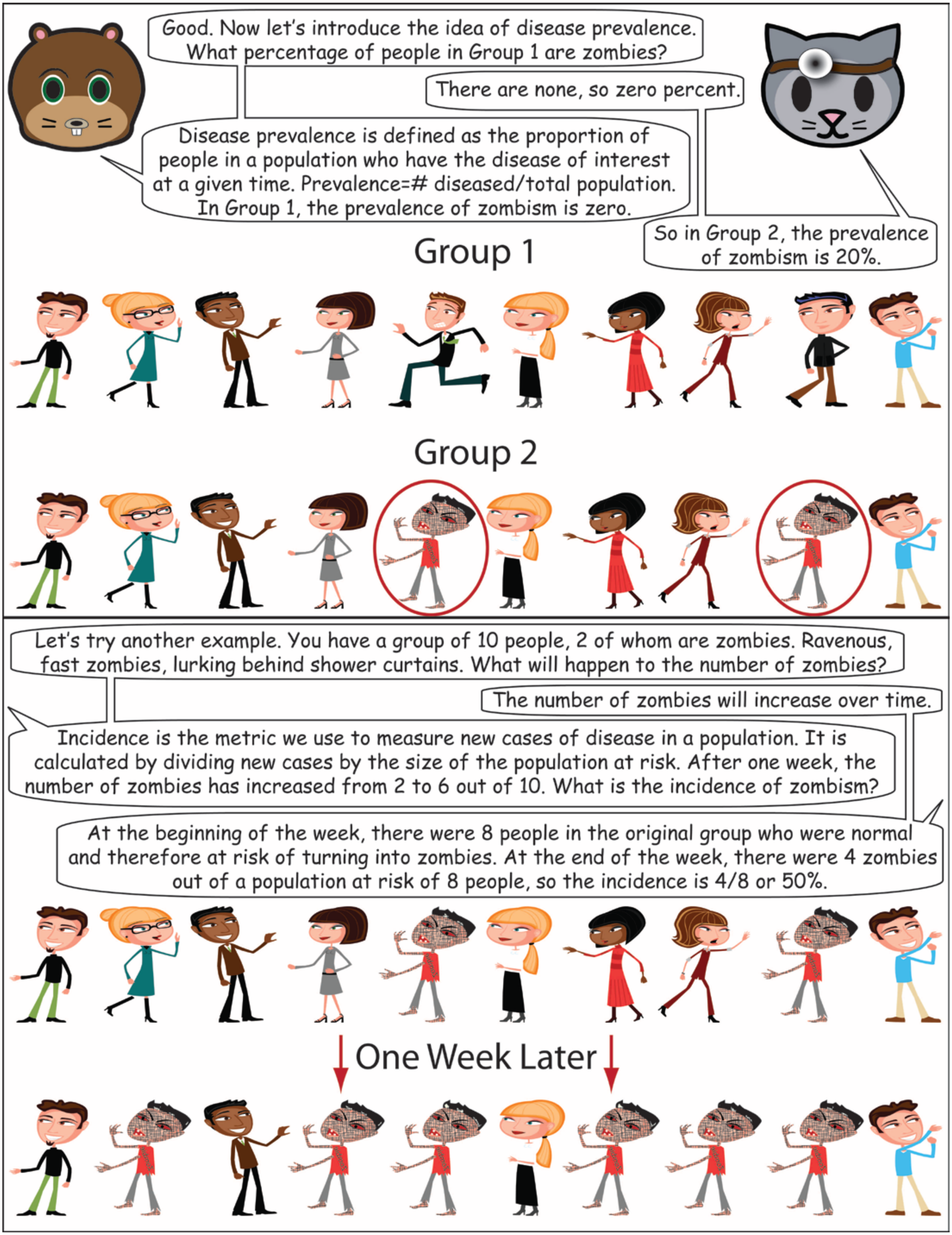
Uh, everybody in London?
Yes. By then, the cholera bacillus had spread to England.
England1832

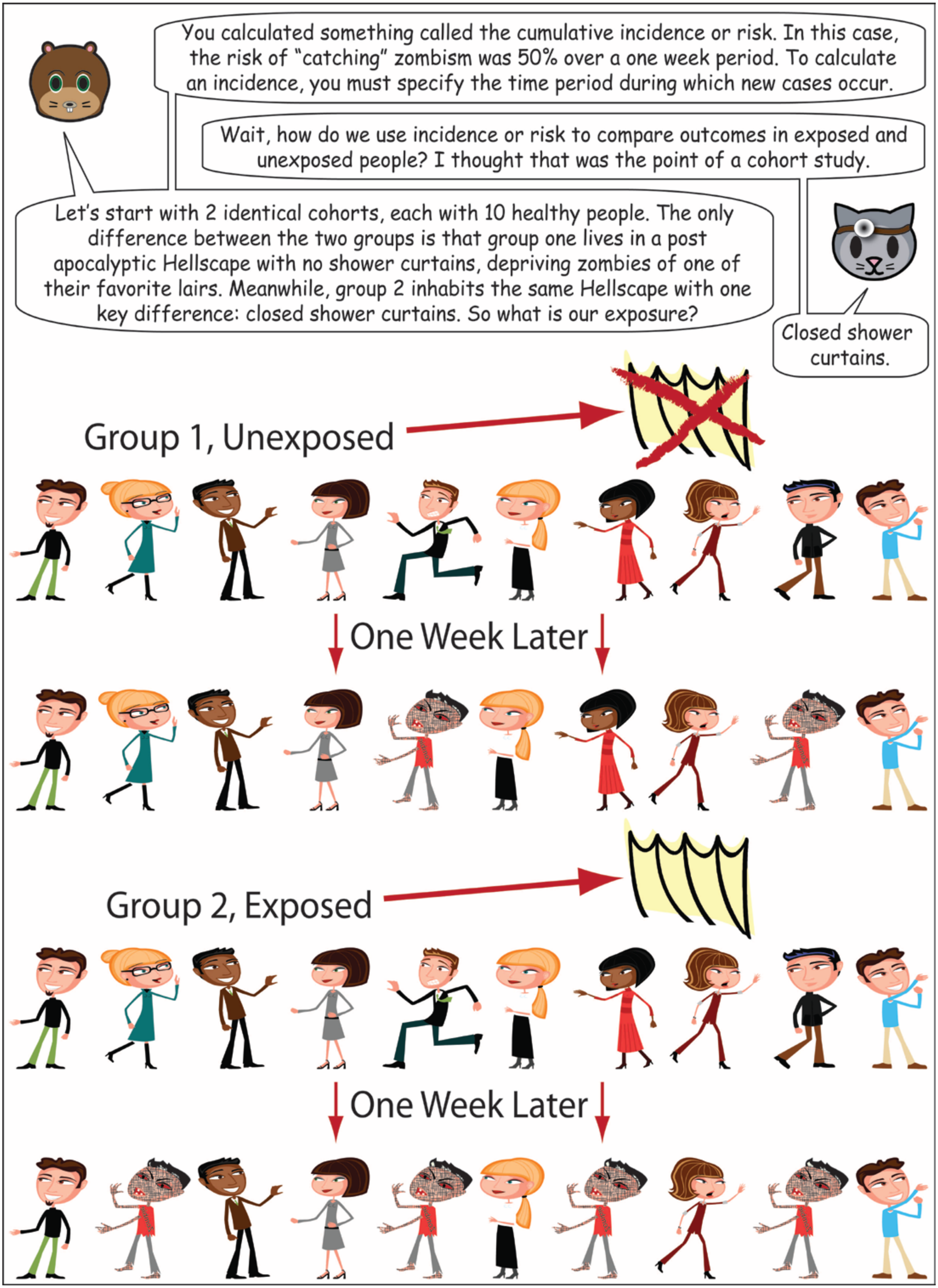

I am still a little confused by population at risk. Do you have any more examples?
If you were studying prostate cancer, you would not enroll any women in your study because women are not at risk for prostate cancer. Enrolling 6 year old boys would be a mistake, since they are also not at risk.

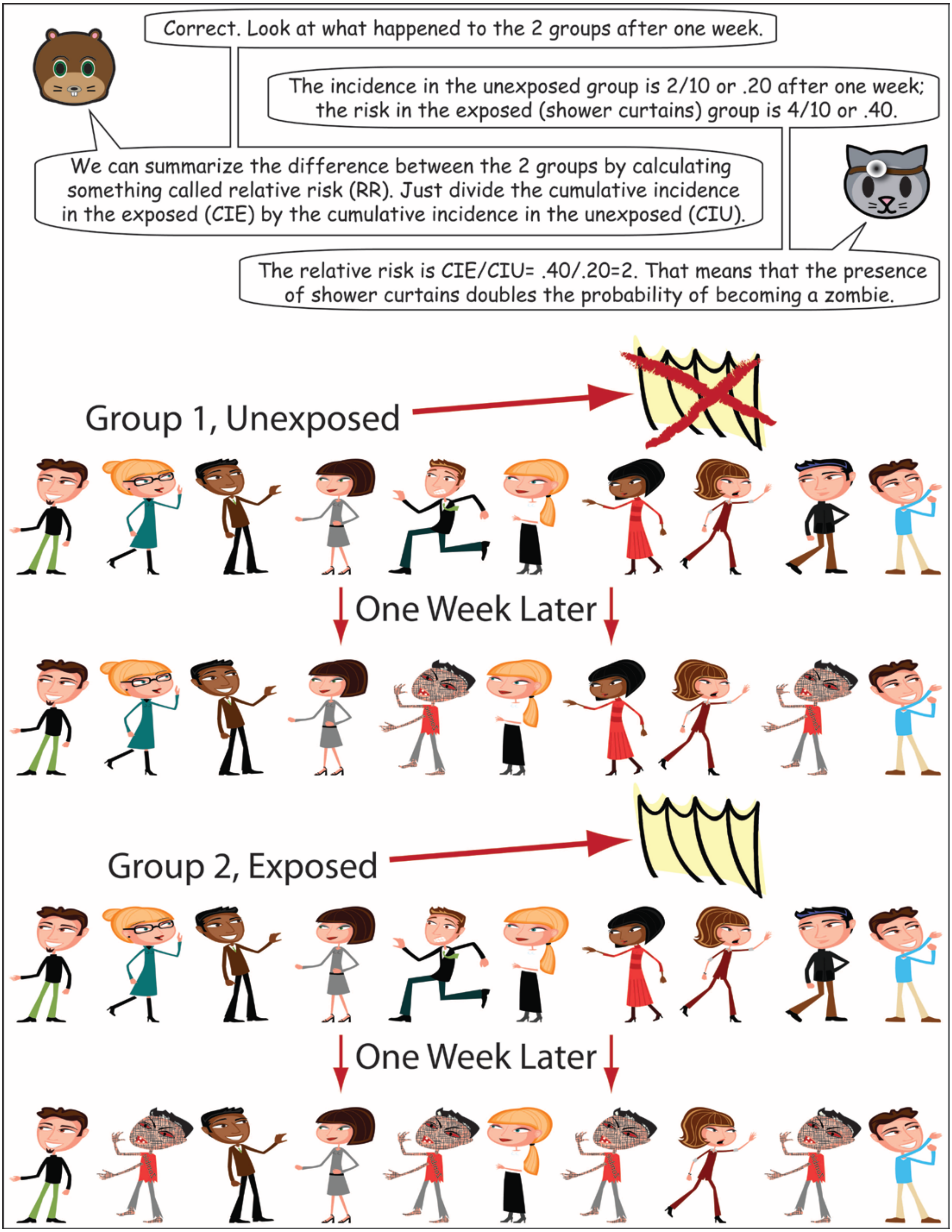

Anything else I need to know about who is at risk?
Yes. If you already have a disease, you are not at risk for developing it. For example, a man who already has prostate cancer is not at risk for developing that disease; he already has it. If you had measles and recovered, you are no longer at risk because you are immune. On the other hand, after you recover from the flu, you go back to being in the population at risk next flu season.

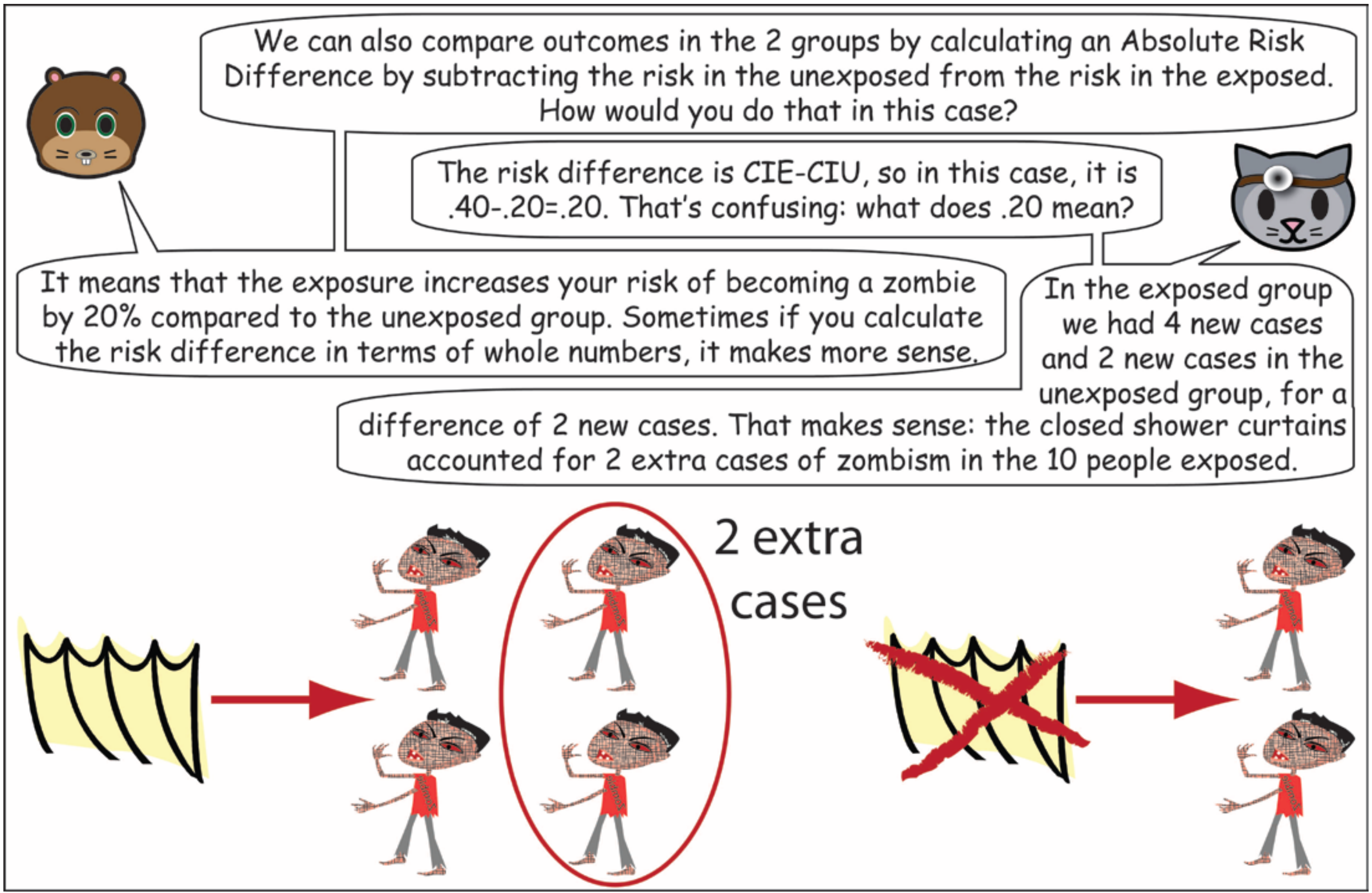
That still seems pretty easy.
True, but we have lots more ground to cover, like incidence, prevalence and calculation of risk. You know what would make this discussion more fun?
A zombie?

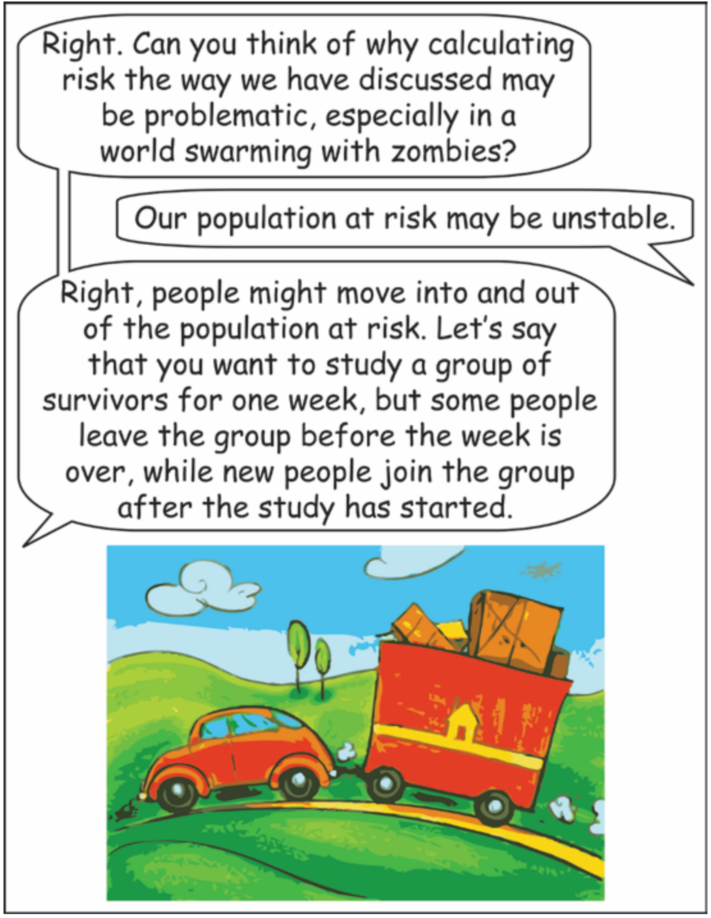

A zombie cohort. Assuming zombism is real, how many people in Group 1 are at risk for becoming zombies?
All 10 of these people.
How about in Group 2?
There are a total of 10 people, but 2 of them are already zombies, so we have 8 people at risk.
Group 1
Group 2

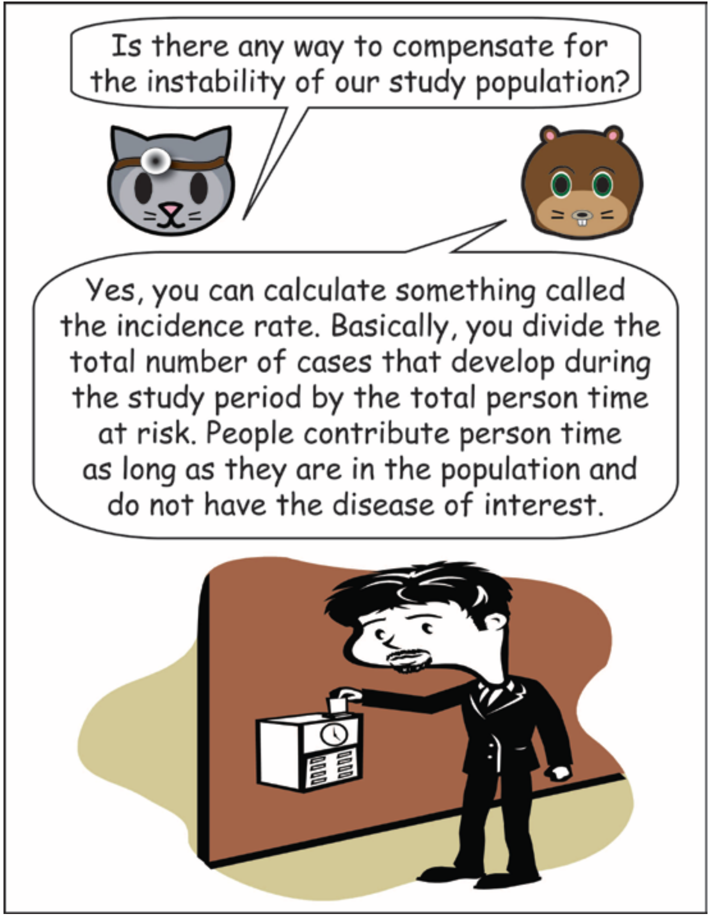

Good. Now let's introduce the idea of disease prevalence. What percentage of people in Group 1 are zombies?
There are none, so zero percent.
Disease prevalence is defined as the proportion of people in a population who have the disease of interest at a given time. Prevalence=# diseased/total population. In Group 1, the prevalence of zombism is zero.
So in Group 2, the prevalence of zombism is 20%.
Group 1
Group 2
Let's try another example. You have a group of 10 people, 2 of whom are zombies. Ravenous, fast zombies, lurking behind shower curtains. What will happen to the number of zombies?
The number of zombies will increase over time.
Incidence is the metric we use to measure new cases of disease in a population. It is calculated by dividing new cases by the size of the population at risk. After one week, the number of zombies has increased from 2 to 6 out of 10. What is the incidence of zombism?
At the beginning of the week, there were 8 people in the original group who were normal and therefore at risk of turning into zombies. At the end of the week, there were 4 zombies out of a population at risk of 8 people, so the incidence is 4/8 or 50%.
One Week Later

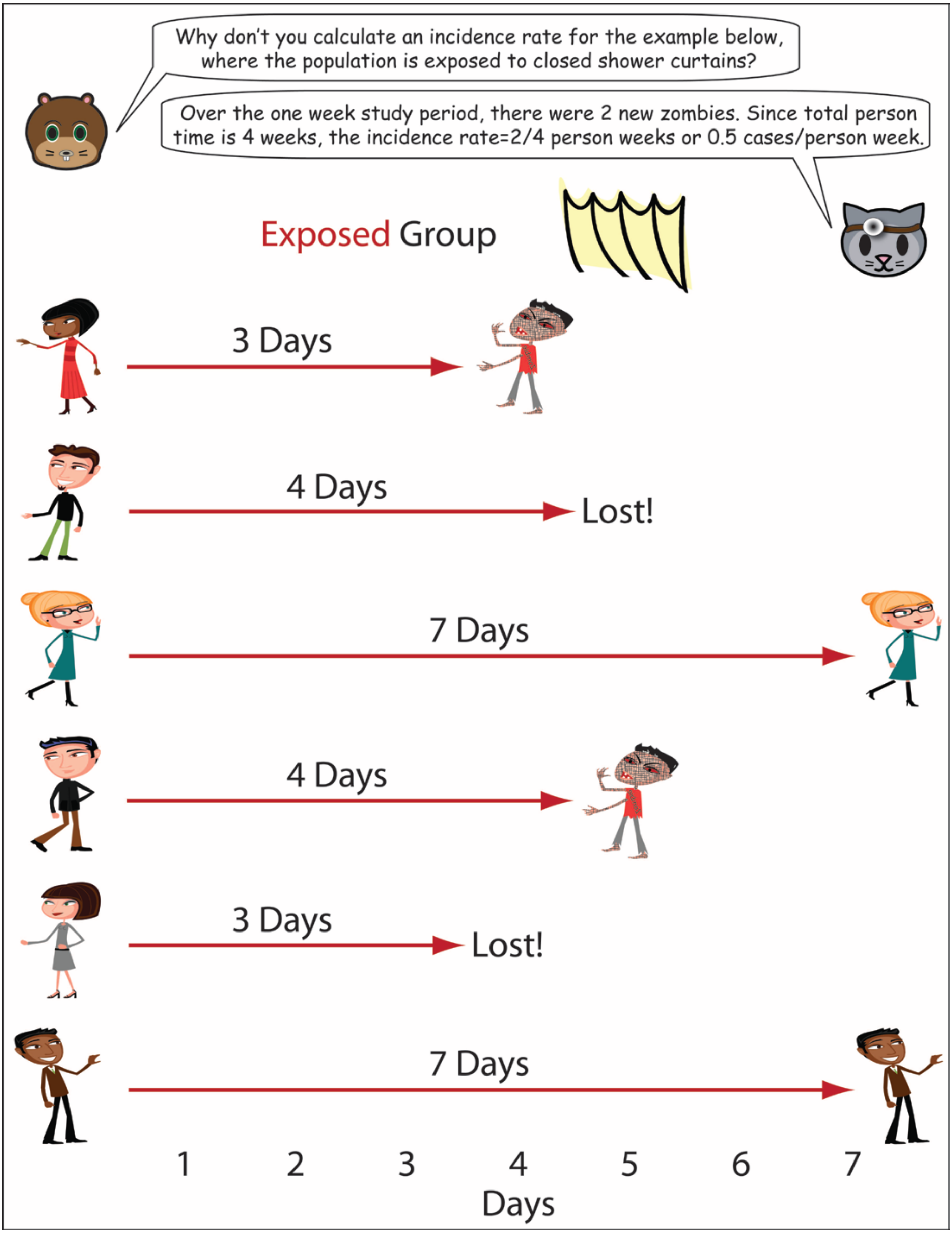

You calculated something called the cumulative incidence or risk. In this case, the risk of "catching" zombism was 50% over a one week period. To calculate an incidence, you must specify the time period during which new cases occur.
Wait, how do we use incidence or risk to compare outcomes in exposed and unexposed people? I thought that was the point of a cohort study.
Let's start with 2 identical cohorts, each with 10 healthy people. The only difference between the two groups is that group one lives in a post apocalyptic Hellscape with no shower curtains, depriving zombies of one of their favorite lairs. Meanwhile, group 2 inhabits the same Hellscape with one key difference: closed shower curtains. So what is our exposure?
Closed shower curtains.
Group 1, Unexposed
One Week Later
Group 2, Exposed
One Week Later

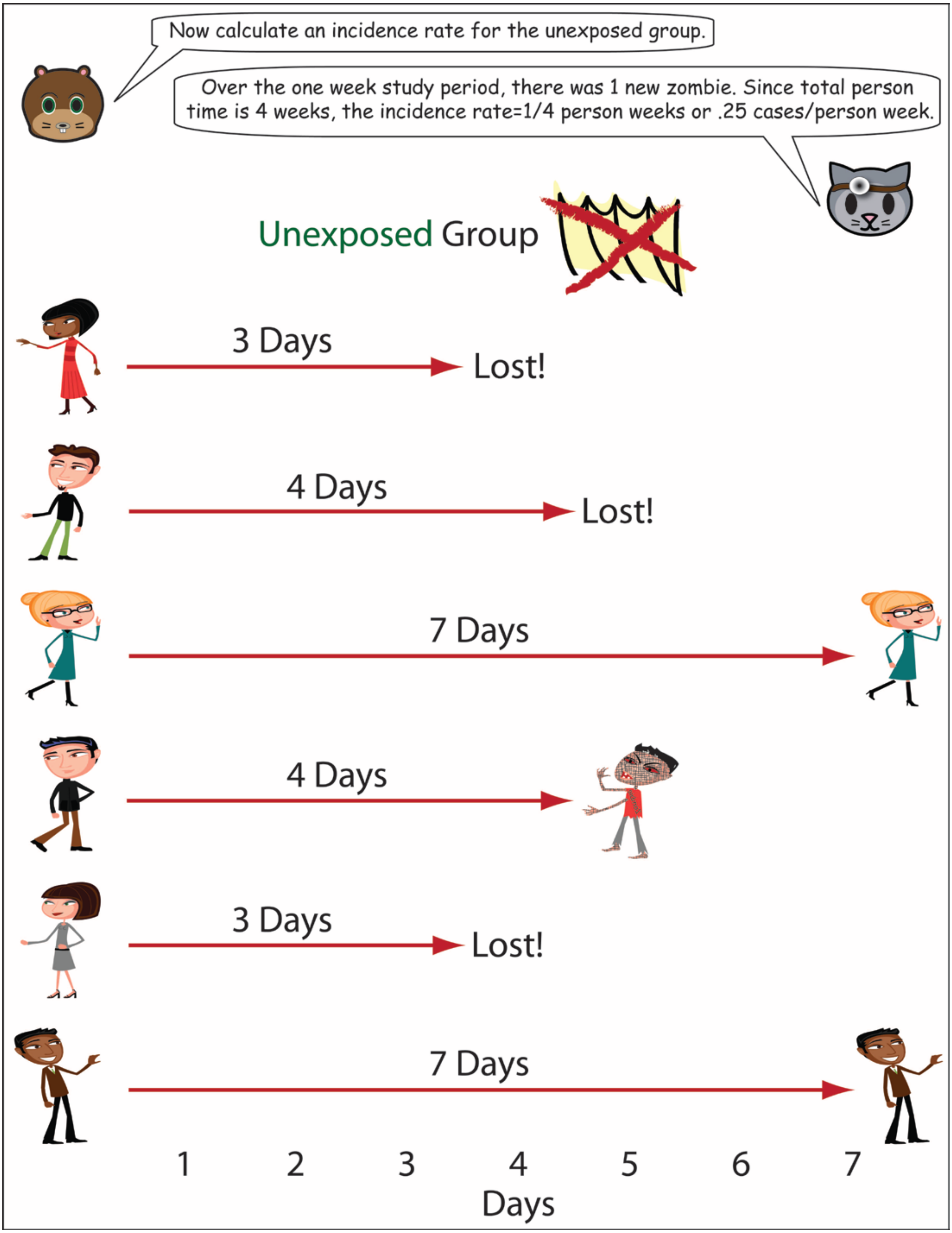

Correct. Look at what happened to the 2 groups after one week.
The incidence in the unexposed group is 2/10 or .20 after one week; the risk in the exposed (shower curtains) group is 4/10 or .40.
We can summarize the difference between the 2 groups by calculating something called relative risk (RR). Just divide the cumulative incidence in the exposed (CIE) by the cumulative incidence in the unexposed (CIU).
The relative risk is CIE/CIU= .40/.20=2. That means that the presence of shower curtains doubles the probability of becoming a zombie.
Group 1, Unexposed
One Week Later
Group 2, Exposed
One Week Later

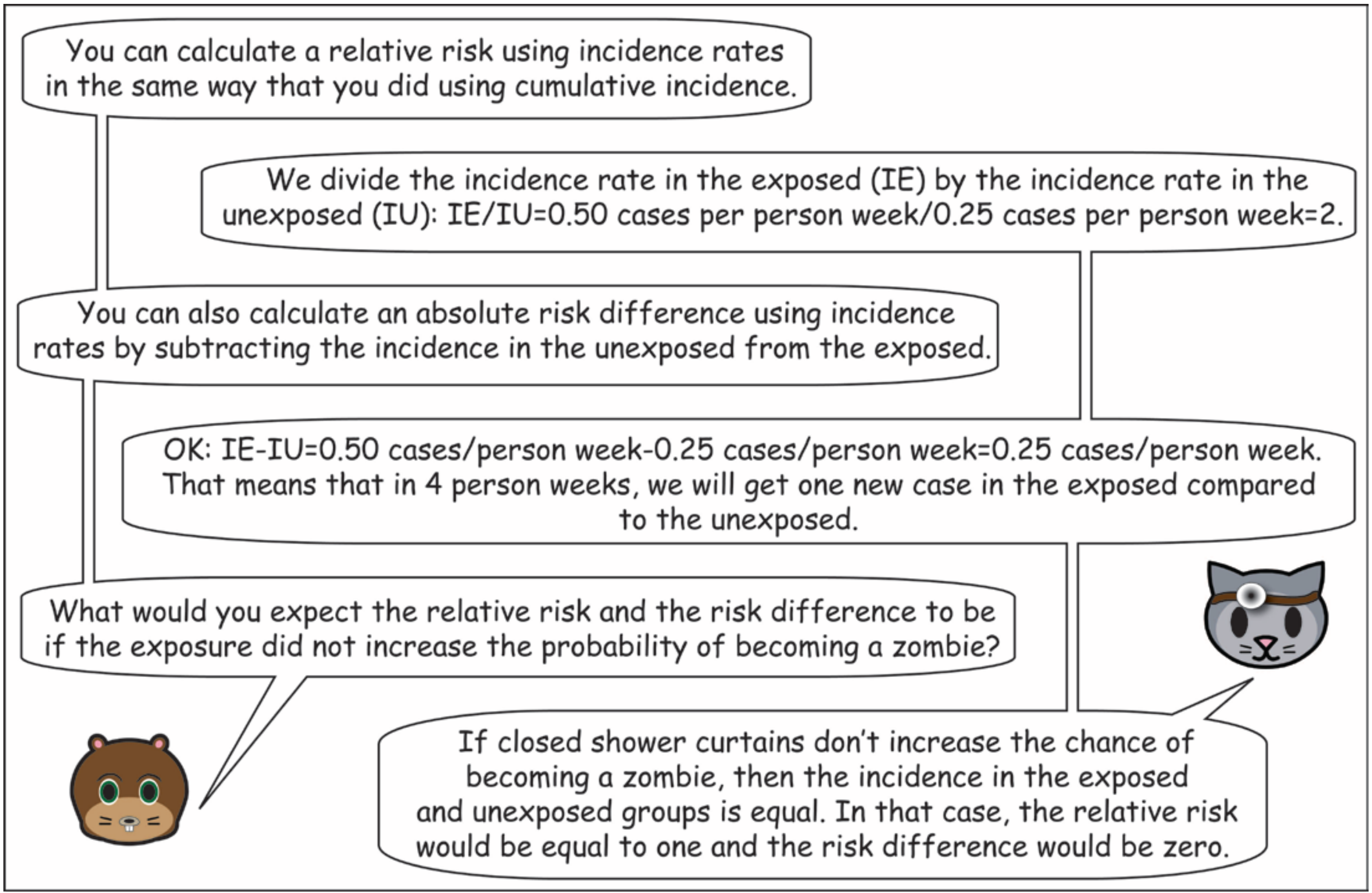

We can also compare outcomes in the 2 groups by calculating an Absolute Risk Difference by subtracting the risk in the unexposed from the risk in the exposed. How would you do that in this case?
The risk difference is CIE-CIU, so in this case, it is .40-.20=.20. That's confusing: what does .20 mean?
It means that the exposure increases your risk of becoming a zombie by 20% compared to the unexposed group. Sometimes if you calculate the risk difference in terms of whole numbers, it makes more sense.
In the exposed group we had 4 new cases and 2 new cases in the unexposed group, for a difference of 2 new cases. That makes sense: the closed shower curtains accounted for 2 extra cases of zombism in the 10 people exposed.
2 extra cases

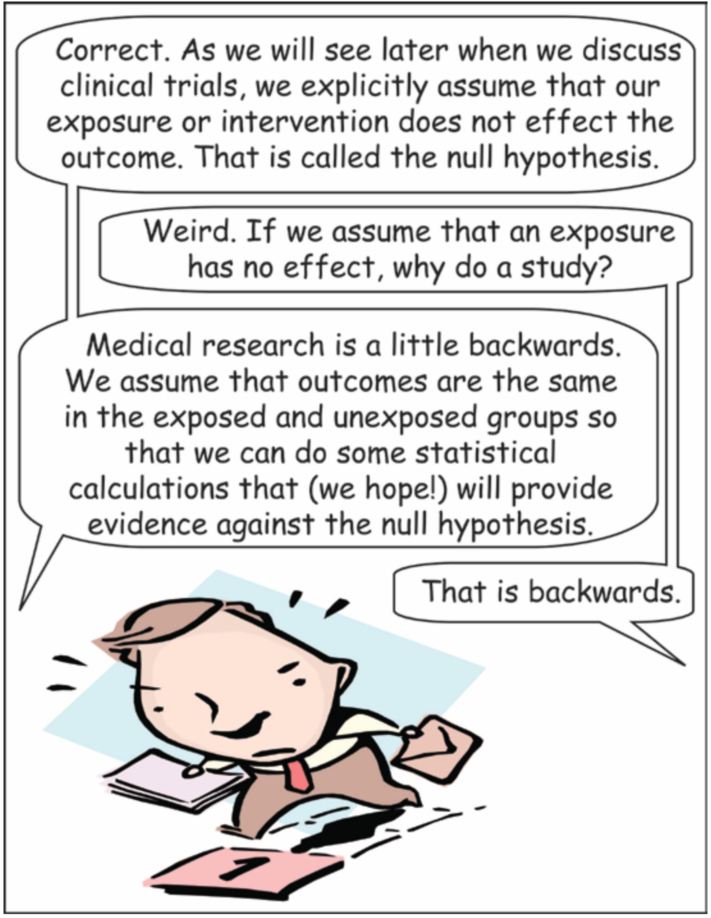

Right. Can you think of why calculating risk the way we have discussed may be problematic, especially in a world swarming with zombies?
Our population at risk may be unstable.
Right, people might move into and out of the population at risk. Let's say that you want to study a group of survivors for one week, but some people leave the group before the week is over, while new people join the group after the study has started.

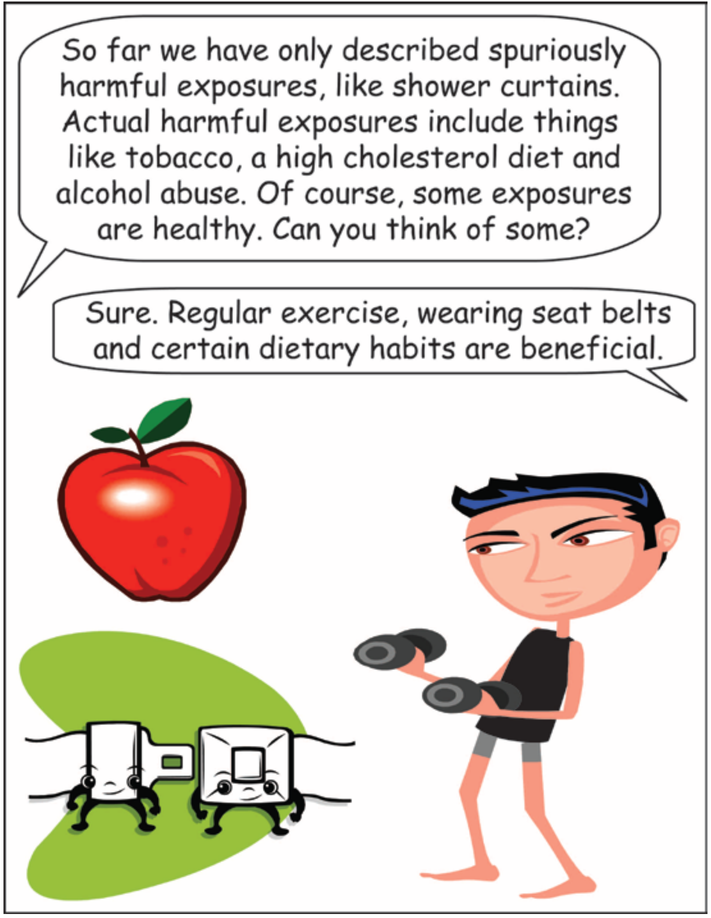

Is there any way to compensate for the instability of our study population?
Yes, you can calculate something called the incidence rate. Basically, you divide the total number of cases that develop during the study period by the total person time at risk. People contribute person time as long as they are in the population and do not have the disease of interest.

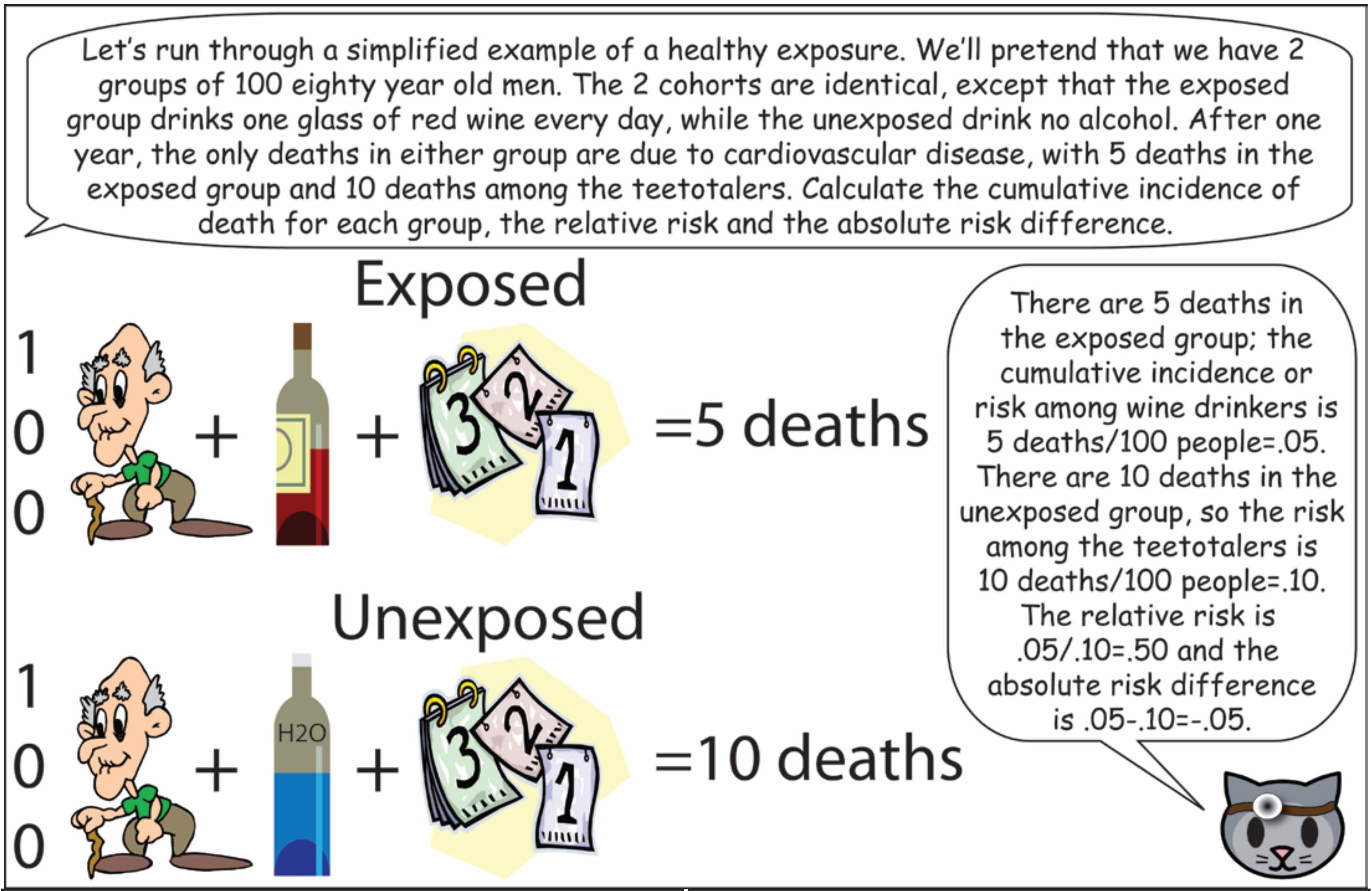

Why don't you calculate an incidence rate for the example below, where the population is exposed to closed shower curtains?
Over the one week study period, there were 2 new zombies. Since total person time is 4 weeks, the incidence rate=2/4 person weeks or 0.5 cases/person week.
Exposed Group
3 Days
4 Days
Lost!
7 Days
4 Days
3 Days
Lost!
7 Days
1
2
3
4
5
6
7
Days

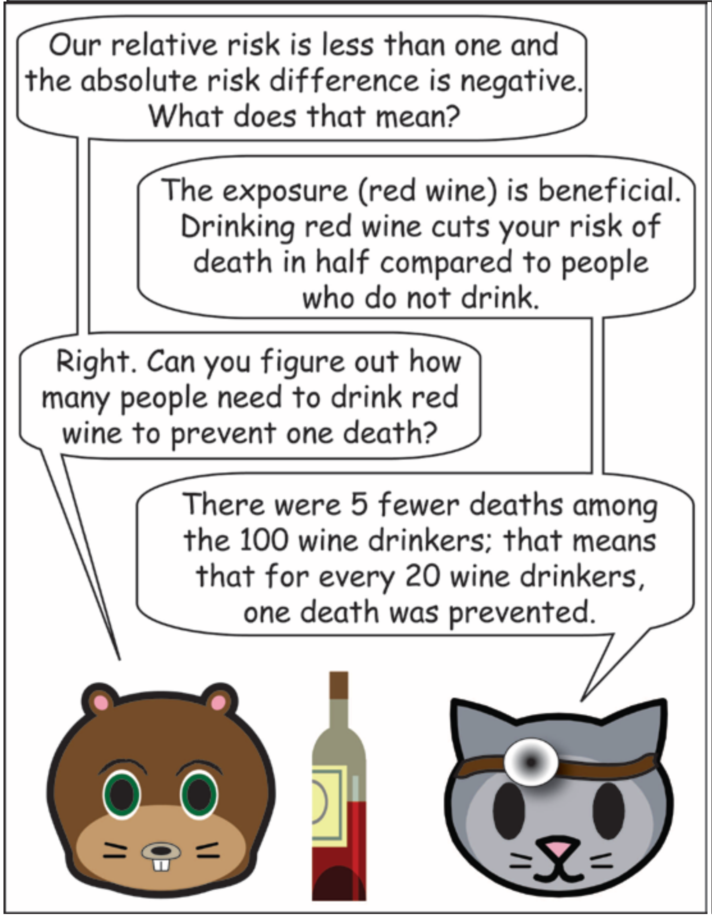

Now calculate an incidence rate for the unexposed group.
Over the one week study period, there was 1 new zombie. Since total person time is 4 weeks, the incidence rate=1/4 person weeks or .25 cases/person week.
Unexposed Group
3 Days
Lost!
4 Days
Lost!
7 Days
4 Days
3 Days
Lost!
7 Days
1
2
3
4
5
6
7
Days

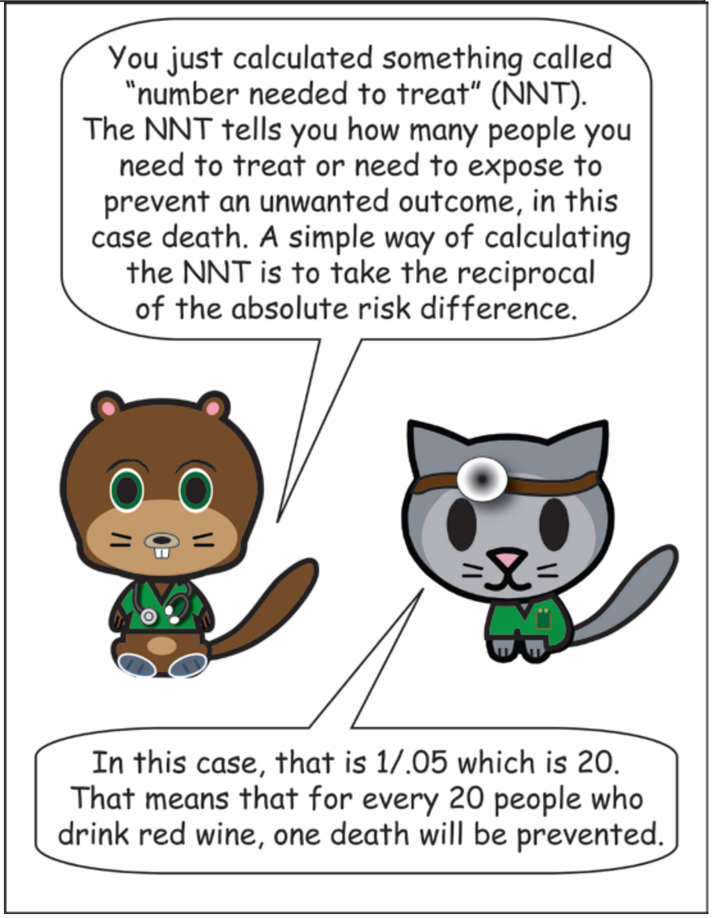

You can calculate a relative risk using incidence rates in the same way that you did using cumulative incidence.
We divide the incidence rate in the exposed (IE) by the incidence rate in the unexposed (IU): IE/IU=0.50 cases per person week/0.25 cases per person week=2.
You can also calculate an absolute risk difference using incidence rates by subtracting the incidence in the unexposed from the exposed.
OK: IE-IU=0.50 cases/person week-0.25 cases/person week=0.25 cases/person week. That means that in 4 person weeks, we will get one new case in the exposed compared to the unexposed.
What would you expect the relative risk and the risk difference to be if the exposure did not increase the probability of becoming a zombie?
If closed shower curtains don't increase the chance of becoming a zombie, then the incidence in the exposed and unexposed groups is equal. In that case, the relative risk would be equal to one and the risk difference would be zero.


Correct. As we will see later when we discuss clinical trials, we explicitly assume that our exposure or intervention does not effect the outcome. That is called the null hypothesis.
Weird. If we assume that an exposure has no effect, why do a study?
Medical research is a little backwards. We assume that outcomes are the same in the exposed and unexposed groups so that we can do some statistical calculations that (we hope!) will provide evidence against the null hypothesis.
That is backwards.
1


So far we have only described spuriously harmful exposures, like shower curtains. Actual harmful exposures include things like tobacco, a high cholesterol diet and alcohol abuse. Of course, some exposures are healthy. Can you think of some?
Sure. Regular exercise, wearing seat belts and certain dietary habits are beneficial.

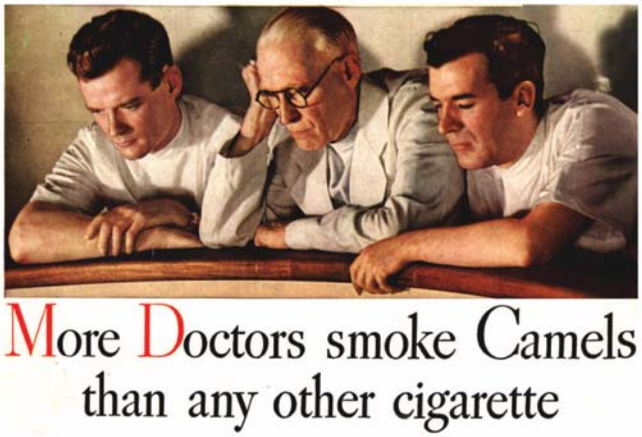

Let's run through a simplified example of a healthy exposure. We'll pretend that we have 2 groups of 100 eighty year old men. The 2 cohorts are identical, except that the exposed group drinks one glass of red wine every day, while the unexposed drink no alcohol. After one year, the only deaths in either group are due to cardiovascular disease, with 5 deaths in the exposed group and 10 deaths among the teetotalers. Calculate the cumulative incidence of death for each group, the relative risk and the absolute risk difference.
Exposed
100 + + =5 deaths
Unexposed
100 + H2O + =10 deaths
There are 5 deaths in the exposed group; the cumulative incidence or risk among wine drinkers is 5 deaths/100 people=.05. There are 10 deaths in the unexposed group, so the risk among the teetotalers is 10 deaths/100 people=.10. The relative risk is .05/.10=.50 and the absolute risk difference is .05-.10=-.05.


Our relative risk is less than one and the absolute risk difference is negative. What does that mean?
The exposure (red wine) is beneficial. Drinking red wine cuts your risk of death in half compared to people who do not drink.
Right. Can you figure out how many people need to drink red wine to prevent one death?
There were 5 fewer deaths among the 100 wine drinkers; that means that for every 20 wine drinkers, one death was prevented.


You just calculated something called "number needed to treat" (NNT). The NNT tells you how many people you need to treat or need to expose to prevent an unwanted outcome, in this case death. A simple way of calculating the NNT is to take the reciprocal of the absolute risk difference.
In this case, that is 1/.05 which is 20. That means that for every 20 people who drink red wine, one death will be prevented.

Good. I think we know enough about cohort studies to evaluate a real one. We will look at a classic; Richard Doll and Austin Bradford Hill's 1956 paper "Lung Cancer and Other Causes of Death in Relation to Smoking". What is the first thing you would do if you wanted to perform a cohort study to determine if there was an association between smoking and lung cancer death?

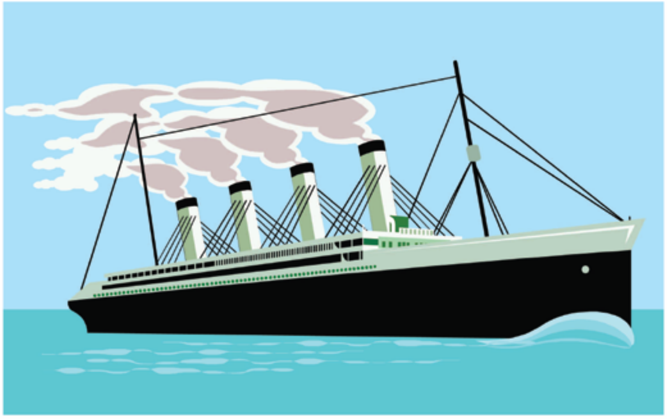
Find a population at risk?



Correct. What population characteristics would make the trial easier to complete?

Since determining the presence or absence of the exposure and the outcome is crucial to the success of a cohort study, we would like to use a group where we can easily establish these things.

Right. You wouldn't choose, say, the crew of a tramp steamer traveling all over the world.

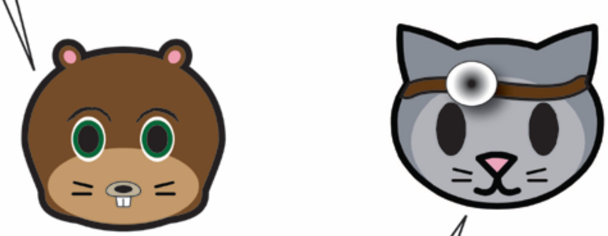


Doll and Hill chose British doctors as their population. All 59,600 doctors in the UK in 1951 were listed in the Medical Register. How do you think Doll and Hill established whether or not the doctors smoked?

By asking them.

LUCKY STRIKE
"IT'S TOASTED"
CIGAR
* The figures quoted have been checked and certified to by LYBRAND, ROSS BROS. AND MONTGOMERY, Accountants and Auditors.
20,679* Physicians
say "LUCKIES are less irritating"

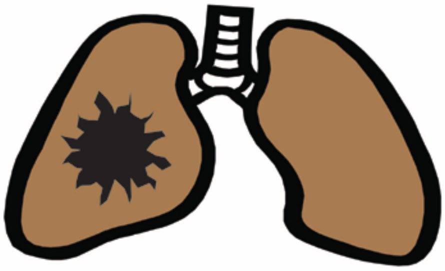
Correct. They sent out a simple "questionary" to all British doctors, asking docs to quantify the amount that they smoked. Respondents were classified as smokers and non-smokers. Smokers were further classified as former, light, moderate or heavy smokers. In a cohort study, it is crucial to correctly establish who is exposed. If an exposure is not socially accepted, it may be underreported; for example drug abusers may deny their habit. If exposure is misclassified, study results are unreliable.



So Doll and Hill compared outcomes between the smokers and non-smokers?

Yes, but we have to deal with multiple other issues first. What people should be excluded from the population at risk?

Doctors who already had lung cancer: they are no longer at risk since they already have the disease. You want to start your cohort study with exposed but otherwise healthy subjects so that you can be sure that the exposure occurs before disease develops.

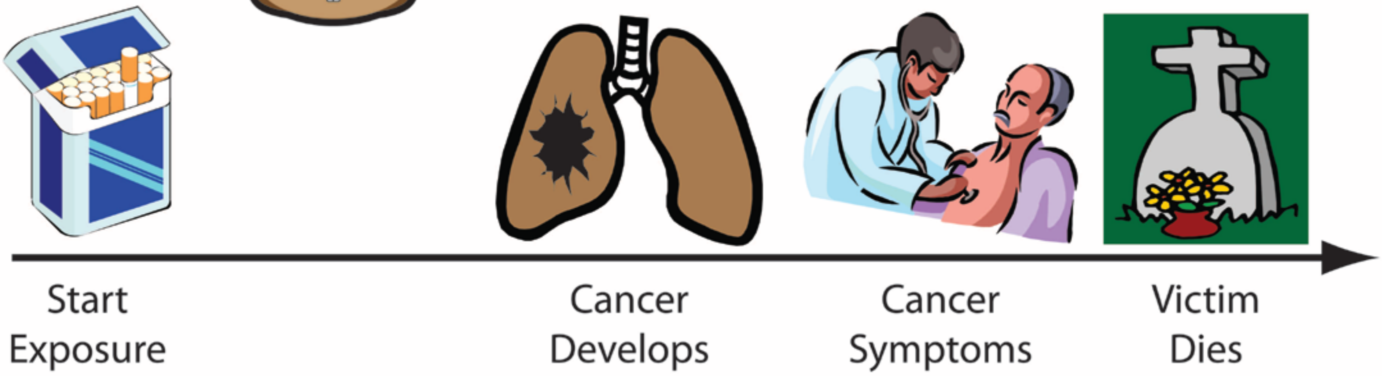


Note: I do not think that they actually did so since they were looking at lung cancer deaths, not lung cancer cases!

True. Doll and Hill also excluded women and male doctors under age 35 from the analysis. Women were excluded because only 2 cases of lung cancer occurred in female doctors, too few for meaningful statistical analysis. Can you guess why males under 35 were excluded?

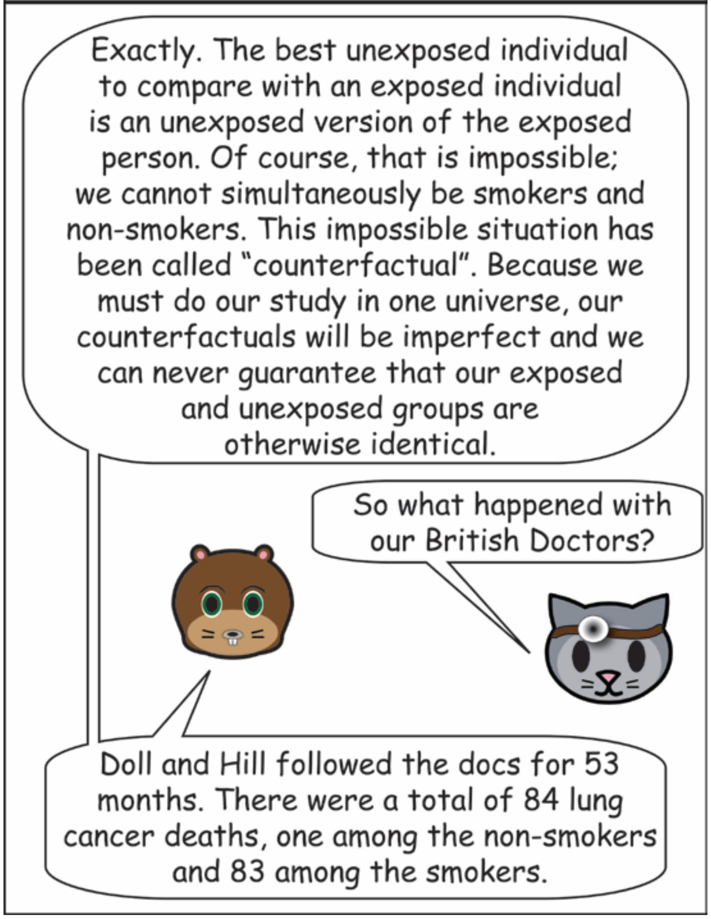
It takes many years of exposure to develop lung cancer.

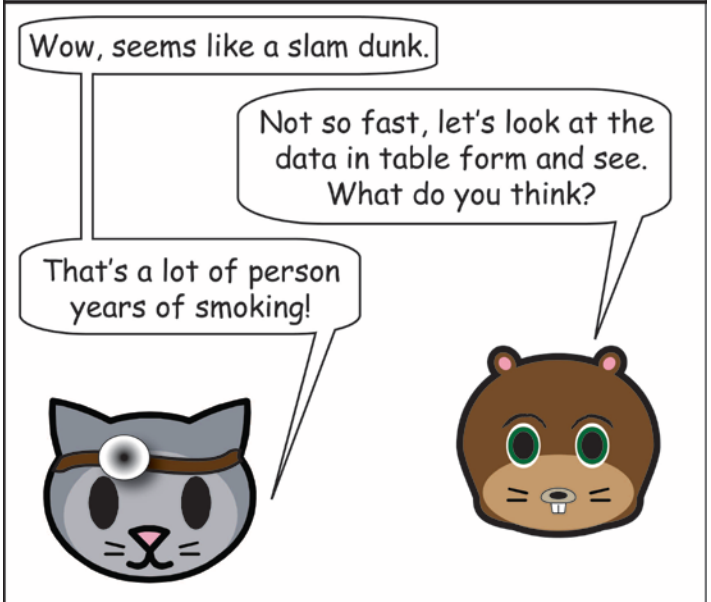
Exactly. It may take 20-30 years of smoking before a lung cancer is detected. This preclinical time can be divided into an induction period and a latent period. The induction period is that time needed to complete the causal bridge that we discussed previously and produce disease. After the induction period is complete and the patient has developed lung cancer, the cancer is initially too small to cause symptoms or be detected by imaging. This portion of the preclinical phase is referred to as the latent period. Doctors under 35 years were still in the preclinical phase, so they were excluded from the study. By the way, the preclinical phase of cholera lasts only a few days, so Snow could safely ignore it when he did his studies.



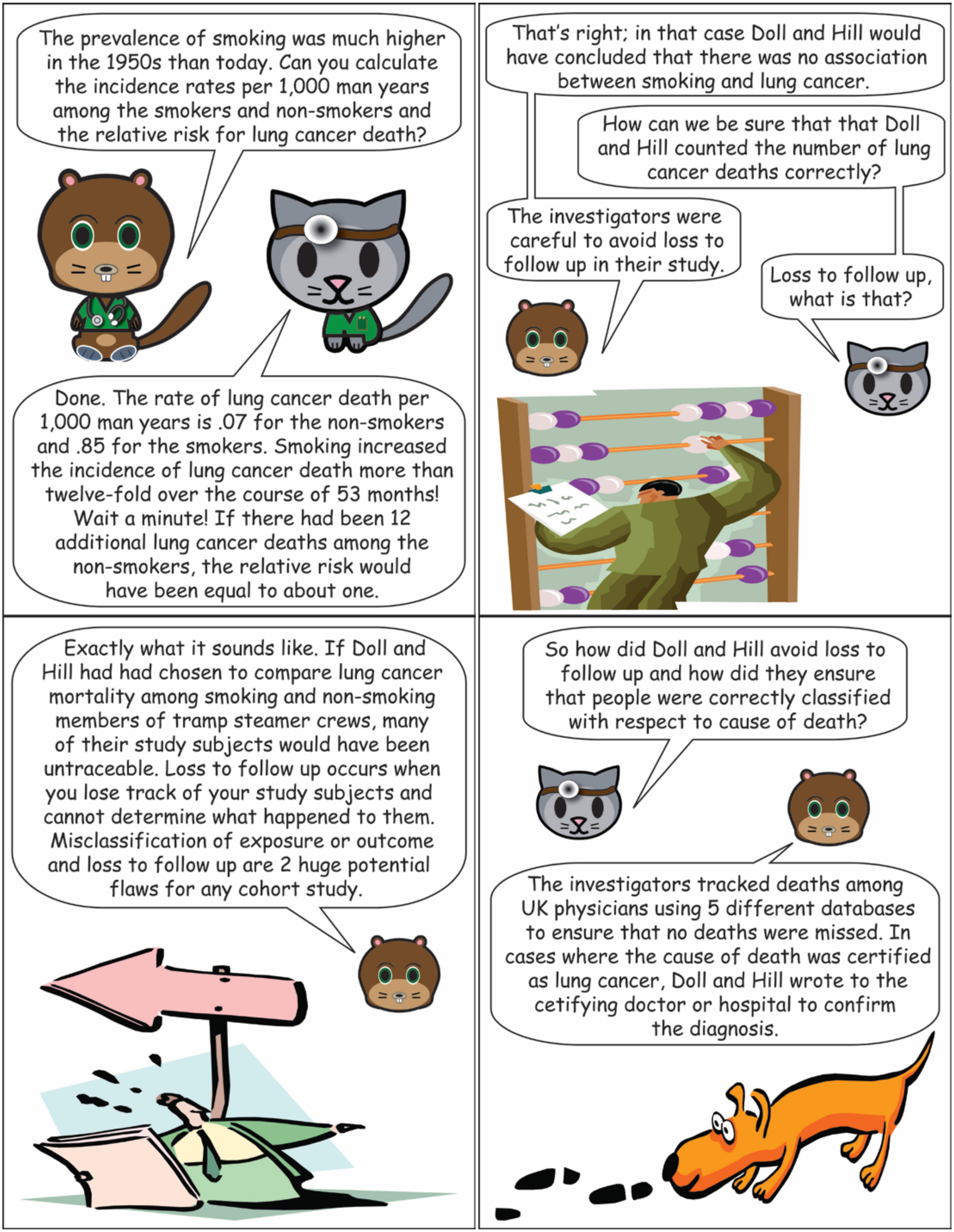
So Doll and Hill compared how often lung cancer deaths occurred in smoking versus non-smoking male British doctors over age 35. But how did they know that smokers and non-smokers were otherwise the same?
Good question. They could not have known for sure, but it seems a reasonable assumption since docs in the UK in 1956 would have been a fairly homogenous ethnic and socioeconomic group. We'll come back to this issue and see how the researchers dealt with one objection to this assumption later on. Unfortunately, a perfect cohort study is impossible; to do one, you would need to be able to travel between parallel universes.

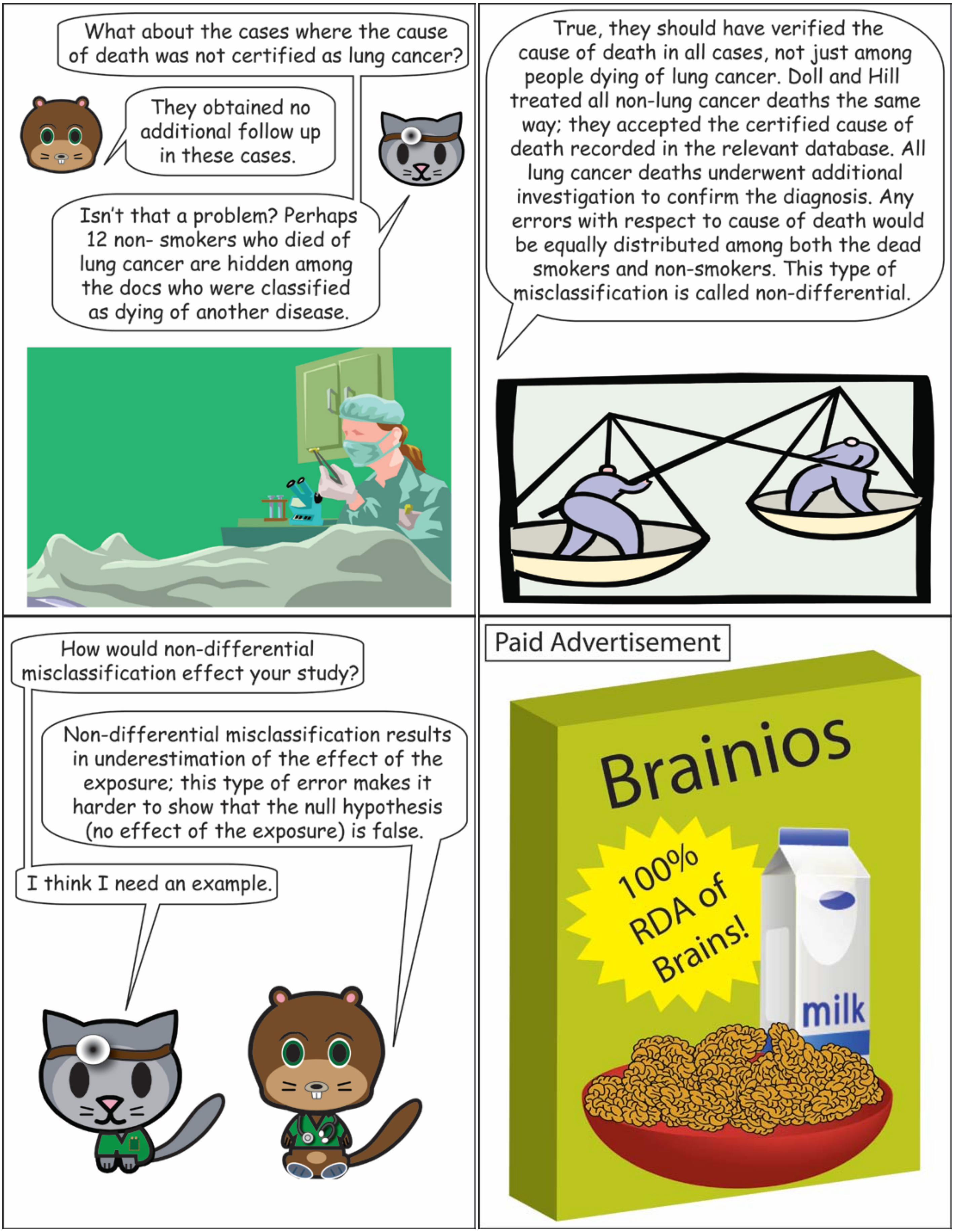
What do you mean?
Let's say I take up smoking on my 40th birthday and you follow me for 30 years to determine the outcome. Who would be the ideal unexposed person to compare with my smoking self?
A version of you that does not smoke.


Exactly. The best unexposed individual to compare with an exposed individual is an unexposed version of the exposed person. Of course, that is impossible; we cannot simultaneously be smokers and non-smokers. This impossible situation has been called "counterfactual". Because we must do our study in one universe, our counterfactuals will be imperfect and we can never guarantee that our exposed and unexposed groups are otherwise identical.
So what happened with our British Doctors?
Doll and Hill followed the docs for 53 months. There were a total of 84 lung cancer deaths, one among the non-smokers and 83 among the smokers.

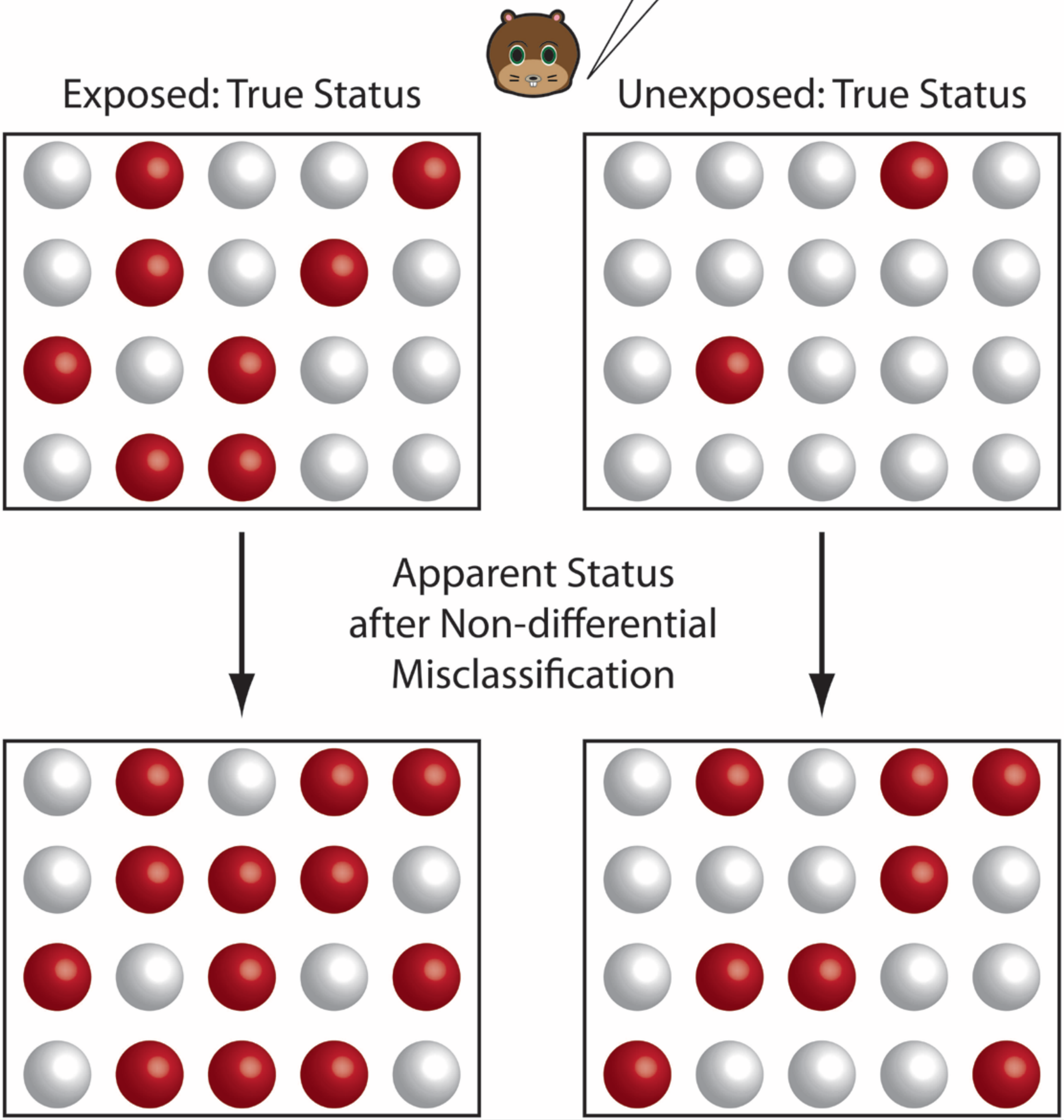

Wow, seems like a slam dunk.
Not so fast, let's look at the data in table form and see. What do you think?
That's a lot of person years of smoking!

	Non-Smokers	Smokers
Lung Cancer Deaths	1	83
Person Years	15,107	98,090


The prevalence of smoking was much higher in the 1950s than today. Can you calculate the incidence rates per 1,000 man years among the smokers and non-smokers and the relative risk for lung cancer death?
Done. The rate of lung cancer death per 1,000 man years is .07 for the non-smokers and .85 for the smokers. Smoking increased the incidence of lung cancer death more than twelve-fold over the course of 53 months! Wait a minute! If there had been 12 additional lung cancer deaths among the non-smokers, the relative risk would have been equal to about one.
That's right; in that case Doll and Hill would have concluded that there was no association between smoking and lung cancer.
How can we be sure that that Doll and Hill counted the number of lung cancer deaths correctly?
The investigators were careful to avoid loss to follow up in their study.
Loss to follow up, what is that?
Exactly what it sounds like. If Doll and Hill had had chosen to compare lung cancer mortality among smoking and non-smoking members of tramp steamer crews, many of their study subjects would have been untraceable. Loss to follow up occurs when you lose track of your study subjects and cannot determine what happened to them. Misclassification of exposure or outcome and loss to follow up are 2 huge potential flaws for any cohort study.
So how did Doll and Hill avoid loss to follow up and how did they ensure that people were correctly classified with respect to cause of death?
The investigators tracked deaths among UK physicians using 5 different databases to ensure that no deaths were missed. In cases where the cause of death was certified as lung cancer, Doll and Hill wrote to the cetifying doctor or hospital to confirm the diagnosis.


What about the cases where the cause of death was not certified as lung cancer?
They obtained no additional follow up in these cases.
Isn't that a problem? Perhaps 12 non- smokers who died of lung cancer are hidden among the docs who were classified as dying of another disease.
True, they should have verified the cause of death in all cases, not just among people dying of lung cancer. Doll and Hill treated all non-lung cancer deaths the same way; they accepted the certified cause of death recorded in the relevant database. All lung cancer deaths underwent additional investigation to confirm the diagnosis. Any errors with respect to cause of death would be equally distributed among both the dead smokers and non-smokers. This type of misclassification is called non-differential.
How would non-differential misclassification effect your study?
Non-differential misclassification results in underestimation of the effect of the exposure; this type of error makes it harder to show that the null hypothesis (no effect of the exposure) is false.
I think I need an example.
Paid Advertisement
Brainios
100% RDA of Brains!
milk

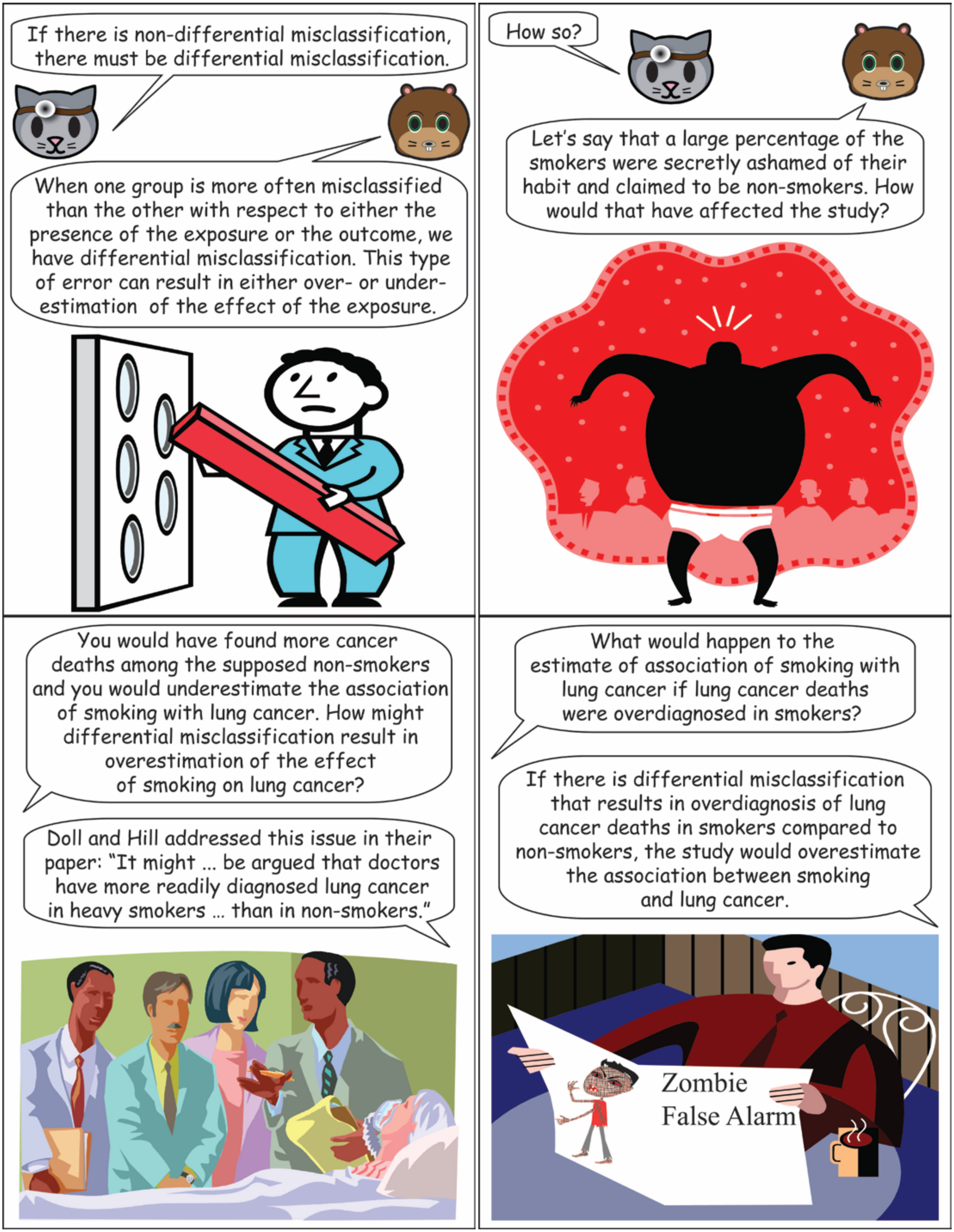
Let's say you are doing a cohort study and have an exposed and an unexposed group. Healthy people are represented by the white balls and diseased people are represented by the red balls. The true health status of each group is shown in the top two boxes: 8 out of 20 exposed are diseased while 2 out of 20 unexposed are diseased. Clearly, there is an association between the exposure and the disease. If there is non-differential misclasification, the association is harder to recognize. The two bottom boxes shows what the groups look like if we mistakenly classify 33% of the healthy people in each group as diseased. With non-differential misclassification, the groups look more alike; now we classify 12 of 20 exposed and 8 of 20 non-exposed as diseased.



Exposed: True Status

Unexposed: True Status

Apparent Status after Non-differential Misclassification

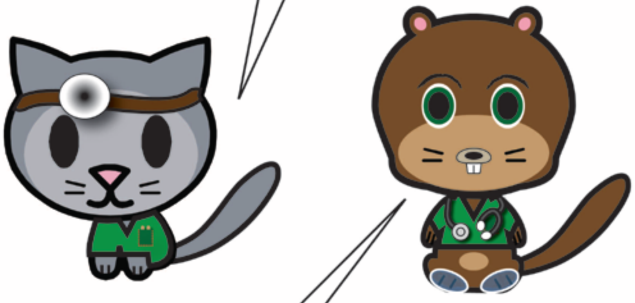

If there is non-differential misclassification, there must be differential misclassification.
When one group is more often misclassified than the other with respect to either the presence of the exposure or the outcome, we have differential misclassification. This type of error can result in either over- or under-estimation of the effect of the exposure.
How so?
Let's say that a large percentage of the smokers were secretly ashamed of their habit and claimed to be non-smokers. How would that have affected the study?
You would have found more cancer deaths among the supposed non-smokers and you would underestimate the association of smoking with lung cancer. How might differential misclassification result in overestimation of the effect of smoking on lung cancer?
Doll and Hill addressed this issue in their paper: "It might ... be argued that doctors have more readily diagnosed lung cancer in heavy smokers ... than in non-smokers."
What would happen to the estimate of association of smoking with lung cancer if lung cancer deaths were overdiagnosed in smokers?
If there is differential misclassification that results in overdiagnosis of lung cancer deaths in smokers compared to non-smokers, the study would overestimate the association between smoking and lung cancer.
Zombie
False Alarm

Doll and Hill presented evidence that showed that there was no differential misclassification that led to overdiagnosis of lung cancer among smokers. Although imperfect, this cohort study provided powerful evidence for the association of smoking with lung cancer. Doll and Hill also found that there was a dose response, with much higher lung cancer mortality among heavier smokers than light or moderate smokers. In addition, ex-smokers had decreased lung cancer deaths compared to present smokers.

That reminds me of Hill's criteria for establishing a causal relationship between an exposure and an outcome. This cohort study demonstrated at least five of his criteria: 1) Strength of Association, 2) Specificity, 3) Temporal Relationship, 4) Biological Gradient and 5) Experiment.

Doll and Hill were incredibly thorough and addressed multiple potential objections to their conclusions. For example, they investigated whether the difference in lung cancer mortality was due to atmospheric pollution and not smoking. Their analysis showed that more non-smokers and fewer heavy smokers lived in big cities, evidence against an atmospheric cause.

Those miasmists are as unkillable as zombies! By the way, how did Doll and Hill ever decide to undertake such a huge, expensive and time-consuming study to examine an exposure that many people thought was harmless? After all, most of the doctors that they studied smoked!



They had evidence from a different type of study published in 1950 that showed that smoking was harmful. This other study is called a case-control study and we will discuss it in the next chapter.

References, Acknowledgements etc.

Many of the illustrations are modified clipart from Microsoft (Redmond, Washington) Office except "Doc" Squirrel is an original creation. All artwork was created or modified using Adobe Illustrator CS4 and/or Photoshop CS4 (San Jose, California). Photographs are all from the public domain. I thank Tom Koch for providing me with a copy of the Oxford cholera map. His book *Disease Maps* is an eye opening, fascinating read for anyone interested in maps and the history of medicine. Look for a chapter on case-control studies coming soon!