Chapter 1

Actuaries Behaving Badly

Let's start by exploring the validity of the most popular pre-post guidelines, as compared to another much less publicized methodology. By then attempting to prove each, we'll learn, in the immortal words of the great philosopher Rick Perry, whether or not who is right.

Some Background: We're Shocked, Shocked, to Find That Invalidity Is Going On in Here

Many people in the care management field complain that nobody believes their numbers. It turns out there's a reason for that: The numbers are mostly made up. I say this with great regret. First, because my job—helping buyers with analysis and procurement—would be much more satisfying if I could deliver good news. Like an East German border guard told the *New York Times* shortly after he was told he could open his gate: "My job is much more fun now that I can let people through instead of shooting them."

Second, the methodology that is at the heart of the fallacies was invented by someone I know well and trust innately—me. Yep, I invented the pre-post methodology I'm about to blow up. Really. You can google "invented disease management" if you like. If you mention pre-post to real biostatisticians, they'll laugh at you. Look in

any public health graduate school course catalog under "biostatistics" and you won't see any reference to pre-post methodologies in the course catalog. The pharmaceutical industry discarded the idea of pre-post as a valid measurement tool almost a century ago, about the same time they stopped experimenting on prisoners. (Ironically, there is at least one disease management program involving prisoners.) The idea of measuring outcomes using a pre-post methodology is so thoroughly discredited in health services research that there wasn't even a book about it until this one, for the simple reason that until the population health improvement field came along, nobody did it.

So, why did population health improvement re-introduce this concept? Two reasons. First, no one—including me, at first—knew any better. Literally, there are only two people prominent in this field with the requisite background for studying population health improvement as though it were an actual academically sound discipline, and at least one of them has pretty much given up on crying in the wilderness.*

Second, when implementing population health improvement programs, few health plans or self-insured employers are willing to deny enrollment to some of their members or employees in order to create a control group getting no intervention, the gold standard for study design. It's not even clear that it is legal under governing law to deny some people an intervention made available to others.

As a result, pre-post study designs covering the entire relevant population have become the norm. Two pre-post methods are in popular use:

- *Prospective Identification*: In this "once chronic, always chronic" methodology favored by many vendors and consultants, a member added to the disease population in any period remains in the disease population in future periods

*Ariel Linden is his name. He has published extensively in this field. He can be reached at alinden@lindenconsulting.org. The other is Ron Goetzel, who will be cited later.

even if the member incurs no disease-related claims in those subsequent periods.

- *Annual Requalification*: Some consultants prefer this methodology, as does The Care Continuum Alliance/CCA. Unlike the prospective identification method, individuals must be re-identified annually through claims data to be counted as disease members in that year.

Both methods embody the consensus of many industry stakeholders. Unfortunately, the consensus pre-post methodology, in all its permutations used to calculate financial outcomes for population-based disease management and wellness programs, lacks the slightest foundation in math.

No matter how strong the consensus, math is not a popularity contest. For instance, the majority of Americans believes that the majority of the 9/11 terrorists were Iraqis. However, none of the 9/11 terrorists were Iraqis. The math trumps the consensus.

During the 15 years in which pre-post designs have been in use by consensus, no one has even attempted to prove the validity of pre-post measurement, which is a good thing because if they tried, they would find exactly the opposite: The attempt to prove that pre-post is valid would result in proving it to be invalid. Nonetheless, most of the industry's vendors and consultants accept the CCA consensus guidelines more or less on faith. And the rest of this chapter is about why that faith is misplaced: The major premise of the CCA guidelines is simply wrong. (The proof will refer occasionally to the CCA Outcomes Guidelines as short-hand for their pre-post methodology. The CCA Outcomes Guidelines cover many methodologies and recommend where possible prospective controlled trials, such as the HealthDialog example in Chapter 7, but cover pre-post in great detail only because that's the methodology most vendors and consultants use.)

The CCA says it does not endorse pre-post (recommending more rigorous methodologies where practical), but rather merely codifies the way consultants and vendors use it. Hence, this section is not a knock against the CCA itself, which claims it is agnostic about the results. They merely provided the forum for a 50-person

Outcomes Guidelines Committee to agree, and occasionally we will use the abbreviation CCA to refer to that group. Since this committee contained no formal hierarchy, everyone's opinions had to be taken into consideration—so perhaps it is possible that the people with the most to lose by valid measurement had the loudest opinions.

How would we know if that's the case? Well, half of me (the half that is convinced there was a second gunman on the grassy knoll) says the CCA knows the major premise is wrong and just put these consensus guidelines out there because they subtly overstate savings most of the time.

The other half of me (the half that thinks, nah, if there had been a conspiracy among the Mafia, the CIA, Lyndon Johnson, and for good measure let's throw in Fidel Castro, someone would have ratted them out by now) says that the CCA is a very well-intentioned organization that recruited the most dedicated volunteers in the industry, who then spent untold hours on conference calls, and tried their hardest to develop an accurate methodology. Unfortunately, the outcomes committee being comprised mostly of people trained in healthcare, they simply missed the mark on the math. An honest mistake.

Which is it—an honest mistake or an attempt to overstate savings? The next edition of the Outcomes Guidelines will answer that question: If the CCA's Committee, having now read this chapter, switches from invalid consensus methodologies to valid proof-based methodologies in the section covering pre-post, then Oswald acted alone.

Why are we picking on the CCA's guidelines? Is it because its authors are dumber than everyone else? No, it's because they're smarter than everyone else, and are the only ones willing to publish their guidelines—free, no less. People who don't use their guidelines would actually be upgrading if they did. If some of the benefits consulting and brokerage firms were to publish their guidelines, a book alone wouldn't do them justice. They'd need a mascot, like "Crime Dog" Fred MacRuff, with a slogan like: "Take a Bite Out of Claims. Let's Reduce Them by More Than 100%."

Want a sense of the level of mathematical sophistication for people who aren't using CCA guidelines? Here's an actual exchange between me and someone who was using individual patients as their own control, an antediluvian methodology whose merits, such as they are, will be extinguished in Chapter 2, a methodology even the CCA Guidelines dismiss. A salesperson was showing my clients and me case studies of her firm's interventions in complex cases, each one resulting in dramatic savings. (It is, as we will see, virtually impossible not to show savings in complex cases—which, as a group, decline in cost after the sentinel events that earned them designation as complex cases take place—intervention or no intervention.)

Case study after case study showing these savings went up on the screen. I asked, politely at first but then with increasing frustration, what the average savings was across these cases. No luck—the case studies continued unabated. Finally, I offered to simply exit the meeting and come back when the presenter was ready to tell us the average. At that point the presenter addressed the issue, albeit in an exasperated tone of voice suggesting that only a complete imbecile could possibly think my question was a good one.

"There is no average," she replied. "It varies."

The Proof: Doing the Pre-Post Math for Grown-Ups

In school, "proof" was a scary word. It also involved a lot of other scary words, like "axiom" and "theorem" and "cosine" and even occasionally "scalene." And, yes, I know we promised that we would only use fifth-grade math in our proofs and case studies, but we lied. It turns out that proving the invalidity of pre-post methodologies requires only *fourth*-grade math.

The validity of both the annual and prospective pre-post methodologies requires identifying the complete "disease population." The fact that many members are not identifiable as "disease members" necessarily creates invalid outcomes, no matter which of the two popular pre-post methodologies is used.

Why Might a Disease Member not be Identifiable to a DM Program in the Baseline?

1. The member's condition is too mild to trigger the algorithm, which normally requires a certain minimum number of prescription fills or diagnosis codes.
2. The member, though diagnosed, is new to the health plan.
3. The member is undiagnosed.
4. The member is noncompliant.
5. The member is misdiagnosed.
6. The member is correctly diagnosed, but as part of a periodic preventive physical, and the physician codes for the physical, not the diagnosis.
7. The member got diagnosed too recently for the claim to have shown up in the data warehouse.
8. The member fills prescriptions using a low-cost generic program, such as Wal-Mart's, and doesn't generate a claim.
9. The member belongs to a culture where having a diagnosis is frowned upon.

We're now going to run a set of tables, simplifying the discussion so that a payor only has two asthmatics.

Consider first the following table. *Prospective identification* (once chronic, always chronic) counts everyone ever identified with the disease from the point of identification going forward. *Annual requalification* counts as disease members only people who trigger the disease algorithm every year. Therefore, Asthmatic #1, who clearly has asthma (having had a $2,000 asthma inpatient event in 2010), gets counted in 2011 only with prospective identification, not annual requalification. The counting strategy yields an obviously incorrect 2011 program outcome for prospective identification. The stated outcome for annual requalification reveals the reality that the year-over-year change in cost is indeed $0.

Person	2010 Costs	2011 Costs	Change in Costs
Asthmatic #1	$2,000	$0 (invisible in annual requalification)	
Asthmatic #2	$0 (Invisible)	$2,000	
Cost per Asthmatic—Prospective Identification Method	*$2,000*	*$1,000*	*−50%*
Cost per Asthmatic—Annual Requalification Method	*$2,000*	*$2,000*	*0*
Actual Cost per Asthmatic	*$1,000*	*$1,000*	*0*

However—and a quirk like this should be a red flag in a demonstration of validity—even though the year-over-year difference is accurately portrayed in annual requalification, the arithmetic average cost per asthmatic itself is overstated by a factor of two ($2,000 versus $1,000).

In another, more common real-world example, assume that Asthmatic #1 decided to control his asthma after his attack by filling $200 worth of prescriptions in 2011. This restates the claims pattern—and distorts the answer—as follows:

Person	2010	2011	Change in Costs
Asthmatic #1	$2,000	$200	
Asthmatic #2	$0 (Invisible)	$2,000	
Cost per Asthmatic—Prospective Identification Method	*$2,000*	*$1,100*	*−45%*
Cost per Asthmatic—Annual Requalification Method	*$2,000*	*$1,100*	*−45%*
Actual Cost per Asthmatic	*$1,000*	*$1,100*	*+10%*

Annual requalification shows a 45 percent reduction in costs per asthmatic, *even as the overall total claims paid climbed 10 percent ($2,000 to $2,200)!* Although this hypothetical could be criticized as an asymmetrical example designed specifically to invalidate the methodology (with more than a modicum of success, I might add), in real life asymmetry is the rule, not the exception: Members are more likely to comply with medication after an attack than before an attack. Therefore, a drug claim is more likely to show up after an attack than before an attack. If the likelihood of an attack was not affected by drug use, making the example symmetrical, the implication would be that drug use would not be valuable in preventing attacks. That is why the more realistic examples would show more drugspend following attacks than preceding them.

Besides, a mathematically valid methodology would work with any claims pattern, symmetrical or not. If something is proven valid, there are no exceptions. Remember the $a^2 + b^2 = c^2$ thing (the Pythagorean Theorem) you learned in sixth grade? It turns out that the equation works with integers only when you square them. There is no set of three integers where you can cube the smaller two integers to sum to the largest one cubed. Further, it turns out that it's not just cubing that's impossible: there is no power and no three integers where raising the two smaller of the three integers to that power equals the third integer raised to that same power. This is Fermat's Last Theorem. It was tested and tested up to extremely high exponents with no exceptions being found. Nonetheless, this Theorem wasn't *proven* until someone showed that no exceptions *could* be found. One exception would have invalidated that Theorem, but now it's a proof.

By contrast, not only do tons of exceptions invalidate the common pre-post methodologies, but it's actually pretty darn hard to find any combination of real-world numbers that support it.

Adding Asthmatic #2 to the disease population will exacerbate the savings misstatement because he gets added only *after* his high-cost sentinel event

Often a member is triggered for entry into the disease population by having an expensive health event. However, adding a person to a cohort once he has presented with a disease—meaning *after* his costs have already increased, or during the year in which his

costs increased—overstates savings. Although the increase in his costs from the prior year doesn't get attributed to the program, his subsequent decline in costs does get attributed to the program. Consider the course of disease in Asthmatic #2:

	2010 (Baseline)	**2011 (Contract Year 1)**	**2012 (Contract Year 2)**
Status	Not in population	Added to population in year of event "trigger"	In population
Asthmatic #2's costs	$0	$2,000	$200
Cost or savings (vs. prior year) attributed to the DM program using either methodology	*N/A*	*$0*	*–$1,800 (savings)*
Summary of Asthmatic #2's impact on savings (2011–2012)			***2011–2012 program savings is increased by $1,800***

Somehow, $1,800 was shown as savings between 2010 and 2012, even though Asthmatic #2's costs per year *increased* by a net amount of $200 from the 2010 to 2012 period. When a vendor or consultant explains, "People are added to the baseline as soon as they are identified with the condition," they mean that there is no mathematical recognition of the baseline experience preceding the sentinel event leading to the identification. To count savings starting with the sentinel event or incident (or the year in which the sentinel incident took place) creates only the illusion of savings. If the accounting shows a year-over-year savings between 2011 and

2012, it must show a loss between 2010 and 2011 in order to capture the reality of the $200 increase in the annual cost for that asthmatic over the full three-year period.*

This fallacy of including people in the measured cohort only after they show up as being high cost or high risk, is even more common in wellness than in disease management. Consider this quotation from a corporate medical director in the *Wall Street Journal* health blog of July 29, 2010: "The same people [lose weight] every year. They get paid for it. They gain weight back. They lose it again. They get paid again." Overstating program success is inevitable.

This dynamic allows some health plans and wellness organizations to claim progress even when there is none. We'll show several examples of this, most notably Health Plan B's wellness program, which has elevated this dynamic to an art form.**

This Vendor A/Health Plan B stuff is very annoying. Can't you simply tell us the real names of the vendors and health plans in the case studies?

Three answers to that question:

1. Sure.
2. Any time.
3. We take Amex, Mastercard, and Visa.

*Most vendors and actuarial consultants prefer to do all calculations in dollars, even though (1) unit cost is not affected by DM or wellness, just the number of units; (2) doing everything in dollars discourages people from checking to see if units actually declined where they were supposed to decline, such as days of care, and increased where they should have increased, such as preventive drug use. I strongly prefer doing everything in units and am writing largely in dollar terms only to "speak the language" of most readers.

** *Why Nobody Believes the Numbers* is about fact, not opinion, so we mean this observation literally, not figuratively. Health Plan A has literally made math into an art form in its bar graphs—see Chapter 4.

Yes, it's true: You need to *pay* for that information (as well as links to the original charts and other information that can't be published due to copyright restrictions) with a separate $29 purchase on www.dismgmt.com. There are several reasons for this. First, in the very unlikely chance that a vendor or carrier changes their offering and measurement to make their results valid, we want to know exactly who received the information about the original invalidity with the vendor's name, so that we can notify them. Second, if there is any other correction to the book needed over time, we can let you know. Third, I have to make some money somewhere. You don't seriously think I'm making any money writing this book, do you? Do the math on the royalties and compare that to the hours and hours of my free time it took to write this thing. I had to forego an entire season of *Survivor*. For all I know a contestant finally got bitten by a snake and I missed it.

Using the Trend of the Non-chronic Population as a Proxy for the Chronic Population will Overstate Savings

The consensus methodologies call for an inflation adjustment. Often, if not always, that adjustment is the trend in the non-chronic population. The CCA, constrained by the voluntary nature of the committee structure to crowdsource everyone's opinions about how they'd like the arithmetic to turn out, is naturally right at the forefront of this scheme, insisting on using at least some of the non-chronic trend with ***bold italics***, just in case anyone has a neurological condition that prevents regular print from penetrating that portion of their brain responsible for making dumb decisions:

> ***CCA recommends the use of a non-chronic population to calculate this trend*** *...* ***defined as those members not identified as having*** *...* ***diabetes, CAD, heart failure, asthma [and/or] COPD.***

The Guidelines then add:

> ***CCA further recommends use of the average historical difference in chronic trend and non-chronic trend to adjust current year trend.****

I'm not sure what the second sentence means, except that the CCA required me to put it in, before they would give permission to reprint the first sentence. Whatever it is, it's not math. Math textbooks don't contain the word "historical." The fifth-grade arithmetic app on your smartphone doesn't include a "compromise" feature that automatically averages two completely different solutions to the same problem. Averaging two solutions reduces your odds of a right answer from a possible 50 percent to a certain 0 percent. It would be like an atheist and a fundamentalist compromising that every other word in the Bible is true.

I don't know what predicts the chronic trend (and in the discussion of valid methodologies, you'll see that you don't have to bother trying), but it's not the non-chronic people who happen to have been findable for some vaguely historical period. As you'll see below, whether you rely on the non-chronic population to supply half your trend or all of your trend or some other random proportion of your trend, and no matter what historical period you look at, it's wrong.

For the non-chronic calculation of trend to be valid, two things must be the case, neither of which is. First, the epidemiology must work: There must be such a thing as a non-chronic population that can be separated from the chronic population. Second, the arithmetic must work out so that, if it is possible to separate the two populations, claims trends in the real world are similar in both populations.

Unfortunately, neither the epidemiology nor the arithmetic withstands the slightest scrutiny.

---

*Care Continuum Alliance, *Outcomes Guidelines* Volume 5, 71, www.carecontinuum.org. In fairness to the CCA, *all* of their recommendations are in boldface italics, not just the dumb ones.

To Paraphrase the Immortal Words of the Great Philosopher Dinah Washington, What a Difference a Trend Makes

Like hanging chads, a slight tweak in trend assumptions could swing the entire financial outcome from positive to negative (or vice-versa).

In this real-life example Table 1.1, note that the savings percentage (3 percent) is far less than the assumed trend factor (21 percent). Building in a 21 percent trend factor changes the calculation from a large loss to a 3 percent gain, meaning

TABLE 1.1 Sensitivity of Savings Calculations to Trend

Baseline Year	All Conditions
Disease-Member Months	150,000
Claims Costs	$50,000,000
Exclusions/Stop Loss (claims in excess of $100,000/person)	$1,300,000
Net Claims Costs (after taking out excluded costs)	$48,700,000
Baseline Per-Disease-Member-Per-Month Costs	$325
Contract Year 3	
Disease-Member Months	169,000
Claims Costs	$68,000,000
Exclusions/Stop Loss (claims in excess of $100,000/person)	$3,200,000
Net Claims Costs (after taking out excluded costs)	$64,800,000
Contract Year 3 Per-Disease-Member-Per-Month Costs	$383
Contractual Inflation Trend Adjustment	21%
Baseline Per-Disease-Member-Per-Month Costs Adjusted for Trend	$394
Baseline Claims Costs Overall, Adjusted for Trend	$64,154,439
Savings	**$2,354,438.65**
Increase in Claims Costs Before Trend Adjustment (%)	**18%**
Reduction in Claims Costs After Trend Adjustment (%)	**3%**
Reduction in Claims Costs After Trend Adjusted ($)	**$11**

(*continued*)

(*continued*)

that a mere three-point decrease in the trend assumption would reduce the savings to zero. And even a one-point decrease in trend assumption—from 21 percent to 20 percent—would reduce the savings by a third (3 percent to 2 percent).

Those with a discerning eye will also notice that the "exclusions/stop-loss" claims removed from the calculation altogether more than double between the two periods. Had there been no high-cost claims exclusion, even with the 21 percent trend there would have been no savings. (Those with a discerning eye may also wonder why claims dollars are rounded to the nearest million while claims savings are calculated to the nearest penny.)

Epidemiology

Let's start with the epidemiology, which requires dividing the population into chronic and non-chronic, and using trend in the latter as a proxy for the former. This fails on two counts:

1. You can't divide a population into chronic versus non-chronic and expect people to stay put in their assigned cohort.
2. The per-patient costs of chronic versus non-chronic migrate differently. You can't use one as a proxy for the other.

So, basically—even before we get to the arithmetic—the whole concept of dividing the population into two defined cohorts falls flat on its face.

The first failure is the fact that people don't stay put: A substantial number of people not identified in the baseline with chronic disease have chronic disease events in the study year.[1] By condition, the percentages are as follows:

TABLE 1.2 Case-Mix of Members Hospitalized in 2005 by Chronic Condition

Category (members)	CAD	Asthma	CHF	Diabetes	Mean
(1) Admitted 2004 year and 2005	5.0%	4.9%	8.8%	11.0%	6.4%
(2) Not Admitted in 2004 and admitted in 2005	69.2%	55.7%	60.8%	36.7%	62.4%
(3) Members *undetected* in 2004 and admitted in 2005	9.0%	14.6%	8.7%	12.7%	10.0%
(4) New Members in 2005 and admitted in 2005	16.8%	24.8%	21.7%	39.6%	21.2%

Categories 3 and 4 represent 31 percent of the population. So the arithmetic assumption that a population can be *a priori* divided into non-disease and disease categories is not just a little wrong, but rather is substantially wrong.

Next, the consensus method claims "stability"[2] between the costs of the chronic and non-chronic populations over time. The correct interpretation, reflecting actual cost trends from 1997 to 2011, is that chronic disease costs in percentage terms rise more slowly than costs in the non-chronic population. This is easy for me to say—I've been writing requests for proposals (RFPs) longer than anyone by far. All I need to do is look at some old RFPs to see that, for example, average annual expense for a heart failure patient in 1998 was about $20,000 at a time when a person without chronic disease cost a health plan roughly $1,600/year—a multiple of 12.5 times. Today, heart failure patients average about $30,000/year while a member of a health plan without one of the five common chronic conditions costs more like $3,200—a multiple of 9.4 times. Though not a huge gap in absolute terms, a difference of this size matters a lot when claimed savings percentages are in the low single digits, like the example in the sidebar above.

I am not the only person to observe this divergence. Another study[3] looked at 16 combinations of four key variables to see whether these trends were similar or not. Those four variables were length of baseline, eligibility period, claims runout, and algorithm used

to trigger identification as having the condition. Each of the four variables had two possibilities. Claims runout, for example, could be three months or six months. If you are keeping score at home and still smarting from being fooled into thinking "5 through 8" represented three chapters, this time we promise no funny business in the following arithmetic: Four variables with two possibilities per variable yield a total of 16 combinations of variables. One of the 16 comparisons showed similar chronic and non-chronic trends. Is the conclusion that there is only one way to design chronic and non-chronic trends to have the latter serve as a proxy for the former? Or should the conclusion be that if one tortures the data long enough it will confess?

The study author's guess is, quite correctly, the latter, because the particular combination of variables that resulted in similar chronic and non-chronic trends had no real theoretical justification other than being one of 16 that he tried. But as a practical matter, even though this author did identify one winning configuration of variables, it makes no difference because I've never once seen an outcomes report with a trend assumption based on that particular configuration.

The drawback of a study involving observations like these two epidemiological studies is that someone can (in the second case) "observe" the other conclusion to suit their bias, or (in the first case) concoct some data to the contrary to suit their bias. As Upton Sinclair said: "It is impossible to prove something to someone whose salary depends on believing the opposite." Or maybe it was Sinclair Lewis. I always get those two mixed up. Like you don't.

Math

So, the first point—the epidemiological one—can be argued. Pathetically, perhaps, but argued nonetheless. Let's then proceed to the second argument, the mathematical one. Fortunately, as we've described, math is not a "he said-she said" discipline. Math is proof-based. In the immortal words of the great philosopher Daniel P. Moynihan: "Everyone is entitled to their own opinions, but not

to their own facts." Something is proven—a fact—if there are no possible exceptions to it, as with the Fermat's Last Theorem example presented earlier. Finding one exception to a proof invalidates it. In the case of the non-chronic-trend-as-proxy-for-trend canard, just like with the pre-post methodology generally, finding an exception doesn't even require breaking a sweat.

Once again, return to the two-asthmatic model. The population is divided at baseline into the disease population and the non-disease population. This division places Asthmatic #2 squarely within the non-disease population during the baseline period because no one at the health plan knows s/he has asthma since no claims have been incurred.* Asthmatic #2 is used as the "what-would-have-been" non-chronic trend for Asthmatic #1.

Person	**2010 (baseline year)**	**2011 (contract year)**
Asthmatic #1	$2,000	$200
—Disease Population		
Asthmatic #2	$0	$2,000
—Non-Disease Population		

Note that the $2,000 cost of an event does not rise between years, meaning that actual unit cost inflation is 0 percent. However, you would never guess that to be the case, if you used the non-chronic trend to estimate the chronic trend. You'd think healthcare inflation was out of control, rising in this admittedly extreme example by an infinite amount. The already-brilliant job of appearing to reduce spending on the disease population by 90 percent ($2,000 down to $200) becomes Nobel Prize-worthy in an environment of infinite healthcare inflation.

*Asthmatic #2 is identified as "healthy" for the purpose of trend calculation because s/he fits one or more of the nine categories listed on page 4 in the "Why Might a Member Not Be Identifiable" box.

Voilà. Using the mathematical axiom that an assertion is invalidated by one example to the contrary, the assertion that non-chronic trend can be used as a proxy for chronic trend is toast. Nonetheless, let's go a step further and create an example to see what happens over time.

In this next three-year table, some members categorized as healthy in the base year (2010) really do have asthma, and all those unidentified asthmatic members have the chance of having an asthma event. No, not every asthmatic's costs will look like #2 in the previous example, but enough do to inflate the non-chronic trend. *This happens because the year-over-year cost increase, up to and including the event, is counted in the what-would-have-been trend calculation.*

If all of us were implanted with transponders that immediately notified our health plan as soon as we crossed a line into having a chronic condition, the pre-post methodology would separate the two populations perfectly. Consider my favorite interviewer, Tim Russert. He was widely assumed during his lifetime to be a non-chronic person with no obvious health problems* and therefore would have been in the non-chronic cohort. His totally unanticipated heart attack therefore incorrectly inflates the non-chronic cohort trend when—with this magic transponder—his event should have been inflating the chronic disease trend.

Yes, we know that unlike most heart attack victims, he died at the scene and therefore incurred no claims expense. *Assuming he had survived*, his claims cost would have been counted in the non-chronic group. Please do us both a favor and try to focus on the bigger picture, okay? Thank you.

Generalizing from his case, absent that magical transponder, *it is inevitable that some chronically ill people will sneak into the non-chronic comparison group and thereby exaggerate the true cost trend of the non-chronic population when they have events.*

Return to the three-year example, and now add a row showing the impact on trend of chronic people being counted as non-chronic:

*Aside from having the world's third-widest head, behind (#2) Alec Baldwin and (#1) Stewey Griffin.

	2010 (Baseline)	2011 (Contract Year 1)	2012 (Contract Year 2)
Status	Not in program	Added to population going forward	In population
Asthmatic #2's costs	$0	$2,000	$200
Change attributed to the DM program using either methodology	*N/A*	*$0*	*–$1,800 (savings)*
Change in cost added to the what-would-have-been trend calculation from previous year		*+$2,000*	
Summary of Asthmatic #2's impact on trend (2010–2011) and savings (2011–2012)		***2010–2011 non-chronic trend is increased by $2,000***	***2011–2012 program savings is increased by $1,800***

The distortion in the disease population's costs over time is exacerbated by the calculation of trend in the non-disease population.

The way you can tell this is happening in reports presented to you is to see if there is a decline in every category of cost, including categories (like drugs and doctor visits) that should be rising in a preventive system. We'll show three examples of this in the case studies.

The reason you know a decline in all categories can't happen is, in reality and in accurately measured programs, claims in pharmacy and primary care always increase if a population health improvement program is successful, as more people substitute preventive drugs

and physician care for hospitalizations and emergency room (ER) care. Our ongoing analogy: insulating your house reduces your energy expense, but your insulation expense rises.

A decline in cost across all claims categories versus trend can be due only to an overstatement of trend, which in turn is a result of putting chronically ill but undiagnosed people in the non-chronic-disease category.

Some of you might be wondering whether the effect of chronic members accidentally assigned to the non-chronic group is offset by the reverse happening, meaning non-chronic members accidentally assigned to the chronically ill group. The answer is, no. In fact, once again, mathematically this mis-assignment would exacerbate the difference in trend between the two groups. It's much rarer for someone to be thought to have (for example) heart disease and not actually have it than the reverse. Hence this next example is a rather unlikely one, but we shall soldier on with it even so. Suppose someone was misdiagnosed with a heart attack in the baseline year, and therefore the $20,000 cost of that person's *faux* heart attack was added into the baseline. Since the person didn't really have heart disease, his costs—which affect the cost of the chronic disease cohort as a group—will likely fall quite precipitously in the following year.

Why does all this happen? How can a well-intentioned committee building a consensus methodology—as well as most of the case studies we will be citing—be so far off? The arithmetic underlying their mistake is quite simple: When they calculate the average costs for a group of individuals with a disease, they don't count individuals with zero costs for that disease* in their average baseline. This causes the average baseline to be overstated, as shown in the first annual requalification example.

Consider a calculation of the average altitude of all the airplanes in the country. By averaging the radar readings, we learn that the average altitude of all the flights in the air (the ones the radar can spot) is 20,000 feet at a point in time. However, the radar has no way of finding planes on the ground. If half the planes in the country are

*In the case of Medicaid, those are often people who are eligible for coverage but who have not enrolled because they have no healthcare costs.

on the ground at any given time, then the average altitude of all the planes is 10,000 feet once the half on the ground is averaged with the half in the air.

A claims extraction algorithm is like that radar, averaging the claims only for people who have enough claims to be noticed, but excluding people who are, in terms of their claims, like the planes on the ground. Excluding individuals with no costs from an average will cause the calculated average to inflate the actual average. Yet, like the planes on the ground, the zero-cost patients—the patients in the "Why a Patient Might Not Be Visible" box—can't be counted because they can't be found.

Recall that the annual requalification method finds no change in the cost per asthmatic.

Person	**2010**	**2011**
Asthmatic #1	$2,000	$0
Asthmatic #2	$0 (invisible)	$2,000
Cost/Asthmatic using Prospective Identification	***$2,000***	***$1,000***
Cost/Asthmatic using Annual Requalification	***$2,000***	***$2,000***

If everybody had that aforementioned magical transponder implanted inside them that beamed a signal to the health plan or vendor the minute their physiology changed from healthy to unhealthy, there would be no invisible asthmatics, no "planes on the ground," ever. In that case, Asthmatic #2 would be counted as an asthmatic in 2010 and Asthmatic #1 would be counted in 2011, yielding the following analysis, which is, of course, the true cost:

Person	**2010**	**2011**
Asthmatic #1	$2,000	$0
Asthmatic #2	$0	$2,000
True Cost/Asthmatic	***$1,000***	***$1,000***

Likewise, in the other example in which Asthmatic #1 incurred $200 in claims in 2011, counting Asthmatic #2 in 2010 reveals that claims rose 10 percent.

Person	2010	2011
Asthmatic #1	$2,000	$200
Asthmatic #2	$0	$2,000
Cost/Asthmatic using Annual Requalification	***$2,000***	***$1,100***
True Cost/Asthmatic	***$1,000***	***$1,100***

Therefore, the two basic tenets on which the entire population health improvement industry is built, pre-post measurement and trending using the non-chronics, are both provably wrong, as a matter of both epidemiology and math. Still, in the interest of fairness, we should let the consulting/vendor industry counter this proof with a counterproof.

The Industry Counterproof to the Epidemiology and the Math

Gotcha! The *counterproof* turns out to be a trick question for two reasons. First, it isn't a question. Second, in math, there is no such thing as a counterproof, which is why if you google the word *counterproof* you get mostly references to engraving. Also, apparently there is a rock group by that name.*

In math, once something is proven, the case is closed because proving the opposite—that is, the aforementioned counterproof—would be impossible. Unfortunately, one of the themes in this book is that many people in the health management industry are unfamiliar with the concept of mathematical impossibility. As John Kenneth Galbraith said, "Faced with a proof that their belief is wrong, 10 percent will accept the proof while the other 90 percent will immediately get to work defending their belief."

*Yes, we agree. That's a dumb name for a rock group. However, rock historians have concluded that most of the good names were used up by about 1985 (along with most of the good songs).

If there were such a thing as a counterproof, defenders of the industry guidelines would claim that they do "adjust" for regression to the mean (all the objections we've been talking about, except the trend issue, fall into the category of regression to the mean). They do so by adding a "lookback" year prior to the baseline, a year in which people who happened to have zero disease-identifiable costs, or be "planes on the ground" during the baseline year itself might have had claims to identify them as having the condition in question. Unfortunately, if you review the list of reasons people with a condition might not be identified in the baseline year, you will see that few people excluded for any of those reasons would be found using a lookback year. To be found through a lookback year but not found in the baseline year, you'd have to have been in the health plan for at least two years *and* have been sick enough to qualify two years ago but not a year ago.

This isn't the only adjustment. The CCA Outcomes Guidelines, following explicit recognition of the limitations of conventional pre-post methodologies, contain page after page of adjustments and alternative methodologies and other caveats to their various methodologies, none of which change the basic problem, which is that the standard methodology is invalid. You can adjust Creationism all you want but it won't result in evolution.

They'd also say that my two-asthmatic example is an extreme one, which of course it is, for the purposes of illustration. Plug any less extreme numbers into those examples, and you'll still get a wrong answer. Not as far off, but wrong nonetheless. That real-life "What a Difference a Trend Makes" sidebar shows that you don't have to be far off—a few percentage points either way totally distorts the underlying result.

That's the epidemiologic problem with their rebuttal. The arithmetic problem with that rebuttal is simple: *An invalid equation cannot be made more valid by adding more numbers to it*. This is amply demonstrated by the immortal words of the great philosopher Captain Louis Renault: "Owing to the seriousness of this crime, I've instructed my men to round up twice the usual number of suspects."

To test that statement (meaning mine), simply go back to all the little asthma tables and substitute a multi-year baseline for a one-year

baseline. Call the baseline however many years you want. Instead of "2010," call it "2006 to 2010." You'll notice that adding years does not create a valid outcome.

What have we learned so far, less than one chapter into *Why Nobody Believes the Numbers*? Quite a bit, it would appear:

1. Like delivering soliloquies, proposing marriage, and cooking broth, math should not be conducted by committee.
2. Trend is invariably going to be measured wrong in pre-post population-based studies—invariably in the direction of overstating the savings in the chronic disease population.
3. There is nothing at all in the realm of either epidemiology or (especially) arithmetic that should lead anyone to use non-chronic trend to predict chronic trend—and yet people do.

Fortunately, there is a way to make lemonade out of the consensus pre-post lemon and turn that methodology into something that mathematicians might recognize as provably valid, and we'll do that next. Epidemiologists and health services researchers will have to wait until Chapter 2 to have their concerns addressed. For now, the solution is to fix the problems in the consensus methodology to create a valid pre-post equation.

Fixing the Problem . . . at Least in Theory

Earlier we noted that the conventional pre-post method would be valid if transponders were implanted in us because then we would know in 2010 who had the disease in 2010, and could put those people in the baseline, whether or not they had claims.

But what if we used a proxy measure? Instead of qualifying people annually, or prospectively, what if we qualified them *retrospectively*, so that once a member shows up as having (for example) asthma in the contract year, we retrospectively include their "zeroes" to recalculate the baseline average claims per member? Ultimately—and it might take a couple of years—all the people who had asthma, diagnosed or not, in 2010 would be populated in the official 2010 asthma baseline.

Start with what is believed at the end of 2010 about the asthma population—there is one asthmatic who cost $2,000:

Person	**2010**
Asthmatic #1	$2,000
Cost/Asthmatic	***$2,000***

The existence of the second asthmatic becomes evident a year later. Populating that asthmatic in the table not just in 2011 (when he is known about) but also in 2010 (when he also had asthma but hadn't been considerate enough to bother telling anyone at the health plan) is exactly what is shown in the last table from the previous section: a valid reflection of actual cost of both asthmatics over both years.

Person	**2010 (re-calculated following 2011)**
Asthmatic #1	$2,000
Asthmatic #2	$0
Cost/Asthmatic	***$1,000***

This valid methodology is called "retrospective identification," as distinguished from the prospective identification and annual requalification methodologies. Here is the table as it looks following 2011, with both asthmatics counted in both years:

Person	**2010**	**2011**
Asthmatic #1	$2,000	$200
Asthmatic #2	$0	$2,000
Cost/Asthmatic using Retrospective Identification	***$1,000***	***$1,100***

It turns out that whereas other methodologies yield correct answers on any given data set about as often as Jupiter aligns with Mars when the moon is in the Seventh House, the retrospective methodology yields correct answers on every given data set. No exceptions. Naturally, being the only mathematically valid population-based methodology in an industry notorious for invalidity, it will come as no surprise that no one currently uses it.

It might also come as a surprise there are two shockingly good reasons not to use it. First, recalculating the baseline every year adds more complexity and uncertainty to a process that for most people is already neither simple nor certain. Second, there is a danger we may *over*-count people with disease. The epidemiology of adding Asthmatic #2's 2010 claims to the baseline once he is revealed as an asthmatic in 2011 is probably sound. He probably did have asthma in 2010. But continually adding people to the baseline once they present with a sentinel event in an "out" year, and then recalculating the baseline to include their claims during that year would be valid only if indeed everyone revealing themselves with a chronic disease in any contract year actually did have the disease in the baseline year. As the contract years accumulate, this approach would overcount people with the disease back in the baseline, adding too many people who really didn't have the condition several years prior to their presentation with it.

The math works every time, but the epidemiology doesn't. It is important nonetheless to show that the math works, to wrap up the discussion of methodologies where the math *doesn't* work to prove that non-working math need not be an integral component of outcomes methodologies. Also, now we have a sound methodology, a methodology that—if we could approximate it in the real world under real-world constraints—would be a useful and reasonably valid tool.

That's precisely where we are going from here: The remainder of this chapter shows how to modify the consensus pre-post formula to approximate the underlying mathematically valid retrospective methodology while still being epidemiologically cogent. Then—because these modifications are observational and not strictly mathematical—the next chapter will show how to test the result via observation, using "plausibility indicators," to satisfy your inner epidemiologist.

$10,000 Reward to Anyone Who Can Prove the Retrospective Methodology Invalid

The author is offering a $10,000 reward to the first industry trade or professional association, outcomes committee, benefits consulting firm, actuarial firm, U.S. citizen, or undocumented alien with a fake Social Security card who proves that, for population-based pre-post analysis, this retrospective qualification methodology is mathematically invalid and that their methodology is valid. Details of the contest are on the www.dismgmt.com website.

Approximating the Valid Methodology in Practice: The Dummy Year Adjustment

Creating a measure that avoids the over/under-counting dilemma requires the use of probabilities. For example, we can't say with more than 25 percent certainty that *exactly* two of four coin flips will be "heads". However, we can say with close to 100 percent certainty that *roughly* 2,000 of 4,000 coin flips will be "heads". Returning to the two-asthmatic illustration helps show how we might apply probabilities to address measuring wellness or disease management efficacy.

Person	**2010**	**2011**
Asthmatic #1	$2,000	$200
Asthmatic #2	$0	$2,000
Cost/Asthmatic using either prospective identification or annual requalification	***$2,000***	***$1,100***

In the absence of any intervention at all, the pre-post methodology generates a whopping 45 percent cost decline. But suppose performing this year-over-year comparison using several different year-pairings—observing what happens in 2010 versus 2009 and 2009 versus 2008—consistently yields a decline similar to 45 percent.

It then becomes possible to compensate for the savings overstated by the invalid pre-post metrics: The first 45 percent of decline would be attributable to the methodology's inherent invalidity, while any further decline would be attributable to the program. The difference between these hypothetical pre-program year-over-year results and subsequent year results could then be used to create an adjustment factor to distinguish program effects from automatic methodology effects.

This factor is called a "Dummy Year Adjustment," and the act of applying it is called a "Dummy Year Analysis." Conveniently, both can be abbreviated as DYA. To return to the coin flip metaphor, consider a situation in which all asthmatics with high enough costs to be identified are "heads" in the baseline year, and 60 percent flip over to "tails" (meaning they become too low-cost to be identified) in the contract year. A typical contractual methodology would credit the vendor with the full 60 percent reduction, but this DYA-based calculation would recognize that 50 percent of heads would flip to "tails" on their own, and credit the vendor only with the additional 10 percent.

DYAs are generated by looking across multiple year-pairings. However, expense, time, or unavailability of data may limit the DYA calculation to two or three year-pairs. One typical *modus operandi* is to analyze two dummy year pairings, and if the calculated year-over-year decline is similar in each one, the average of the two declines becomes the DYA. If the year-pairing declines are dissimilar, a third and even fourth year-pairing is undertaken in order to hone in on the decline due to methodology.

Wait a Second—Aren't Those Different People in Each Year-Pairing?

Of course. In DM you're always talking about different people in each year. But it's the same condition, the same algorithm, the same organization. Especially if you run multiple dummy year-pairings, the people involved in each year-pairing should have similar characteristics as a whole, even if they are not the same exact people.

It is hard to imagine the Illinois state government producing a "Gallant" example of anything involving moral or financial integrity, being a "Goofus"* type of state in those two respects.** Nonetheless, an excellent real-world example of a DYA would be Illinois Medicaid's "frequent flyer" emergency room diversion program.

The state's Medicaid agency wanted to identify and educate high ER utilizers about using alternatives to the ER, and then measure to see if that education made a difference in their subsequent ER utilization. Instead of just identifying everyone who had five or more visits in a year, educating them, and seeing how many visits they made the subsequent year, they started by tracking the subsequent year's performance for high utilizers *before* a program was put in place. It turned out that as a group, people with five or more visits in a single year went to the ER 40 percent less in the subsequent year even without a program, meaning that one year's highest ER utilizers were not necessarily the next year's highest utilizers. This regression-to-the-mean decline proved remarkably consistent over five retrospective year-pairings.

So instead of crediting the vendor with gross reductions in utilization of the ER by the identified high utilizers in the baseline, the state credited the vendor only with reductions beyond the automatic 40 percent heads-to-tails effect. This resulted in the vendor showing modest improvements rather than the massive savings they would have taken the credit for otherwise.

Postscript: Illinois can't stay out of character for long and will make a cameo in the "100% Club" later on in this book, enthusiastically joining the list of states claiming more savings than the amount they spend on chronic disease events, thus violating the rule in math that you can't reduce a number by more than 100 percent no matter how hard you try. And rules in math are so strictly enforced that even Rod Blagojevich can't violate them.

---

*Gallant politely reminds readers that both he and Goofus are registered trademarks of *Highlights for Children*. Goofus sprinkles Gallant's DNA at a crime scene.

**Statistically speaking, you have a better chance of going to jail in Illinois by becoming governor than if you kill someone.

Some Do's and Don'ts for Dummy Year Adjustments

The DYA fixes the annual requalification methodology by canceling out its invalidity. (Prospective identification is too invalid to be fixed by anything.) However, a methodology has to be applied consistently year over year in order for the DYA to yield similar results in each year-pairing. A year with a lot of outliers—a year in which a significant benefits design change took place, or a year with a significant demographic change—will skew the results on either end of the year-pairing. And for that matter, it would skew other methodologies even more than they are already skewed.

For instance, a dramatic reduction in co-pays for preventive drugs or a dramatic increase in co-pays for ER visits might be enough of a change to bend the event rates with or without a care management program. Likewise, although steady aging in a population will not affect the DYA calculation, a layoff, early retirement incentive, or merger will prevent a consistent result. These confounding variables make it impossible to attribute or even correlate an outcome with a program no matter what methodology is used.

Confounding variables do not necessarily undermine the usefulness of the DYA. The benefits design change has to be a substantial one in order to throw off the calculation, since any design change would have to strongly discourage or encourage preventive or curative care relative to the previous year. Finally, the "plausibility test" discussed later will be able to approximate the impact of these confounding variables.

As with any methodology and not surprisingly quite the contrary to most other guidelines, it is preferable to use the DYA on units of utilization, rather than on unit costs. Adding unit costs into the equation adds the likelihood of mistakes due to variations in inflation trend. As noted earlier, using the non-chronic trend to estimate chronic trend will usually, if not always, overstate actual inflation. And as noted in the example sidebar, the magnitude of the trend adjustment tends to overwhelm the magnitude of the savings.

The problem in using dollars for a DYA is that actual inflation itself, even if validly calculated, also varies by year. Also, because unit cost contracting is not involved in any care management initiative, the likelihood that using cost-based metrics will add any insight is overwhelmed by the likelihood of distortion. The focus of disease

management is on reducing units of care, not the contracted cost per unit of care. Hence, there is no analytical reason to try to factor the latter into the calculation.

The DYA factor is accurate only if the previous program (the one in place during the year-pairings) was ineffective or nonexistent. In those cases, the entire year-over-year reduction would be due to the methodology. The DYA loses its usefulness when the program in place during the dummy year-pairings was equally effective during all the years in the pairings, meaning that program effectiveness could be a confounding variable for the methodology overstatement in the year-pairings. Suppose, for example, that a DYA consistently shows, for example, a 5 percent savings in different year-pairings. How can we know much (if any) of that 5 percent is due to the previous program rather than the methodology?

The next chapter answers that question using an observational analysis based on event rate measurement called a "plausibility test." In almost any population health improvement program, the savings can only come from reducing adverse events (or some other easily trackable resource, like specialist visits). If the plausibility test reveals no changes in adverse event rates, then the year-over-year reductions shown by a pre-post—even with a DYA—will be due to regression to the mean. If, however, the plausibility test shows that an organization has enjoyed a reduction in event rates from previous years, the year-over-year cost improvement was due at least in part to event avoidance.

Let's close this chapter with one takeaway that binds the math together, a takeaway that is quite the opposite of most other guidelines, and one that can be applied generally to life. We, too, will use bold italics, accepting the risk that the CCA may sue us for font infringement:

> ***Test multiple methodologies on a simple hypothetical where the right answer is obvious to the naked eye. The methodology that yields that naked-eye answer is the right methodology. All further refinements should be applied to that methodology (or an approximation of that methodology), as no amount of refining can turn invalid methodologies into valid ones.***