

**DRUG SAFETY AND DEADLINES:
AN APPLICATION OF PROGNOSTIC SCORES, PROPENSITY SCORES, FULL
MATCHING AND PERMUTATION INFERENCE TO REGULATORY REGIME CHANGE AT
THE FDA ¹**

Jake Bowers, Dan Carpenter, Ben Hansen

May 10, 2007

Abstract

This paper addresses two questions, one substantive and one methodological: The substantive question asks, “Have the imposition of deadlines on FDA review by the Prescription User Fee Acts (PDUFA)s caused drug safety problems?” The methodological question asks, “In an observational study with discontinuities, how can we use as much information as possible (in order to enhance precision) while avoiding bias and fragility of results as much as possible?”

We propose to answer the second question using propensity scores, prognosis scores, full matching, and permutation inference (as well as discontinuities in the FDA regulatory regime). We develop our answer to the first question throughout the paper.

1 PDUFA

In September of 1992, the FDA changed from an agency responsible for the safety of drugs sold in the USA to an agency responsible for *both* drug safety *and* speedy drug approval. In order to help the FDA meet the new deadlines, the 1992 Prescription Drug User Fee Act (PDUFA)² changed the FDA such that (1) the FDA budget was increased by virtue of fees directly paid to it by drug companies and (2) specified that the FDA was to use this extra money to hire many more reviewers (which it did) in order to come to decisions about drugs much more quickly than before (which it did).³ Advocates of the act have based their support on the idea that that approving drugs more quickly saves lives. Opponents of the act have based their opposition on two claims: (1) regulators should not have their budgets depend directly on regulatees and (2) drug safety is jeopardized by hasty drug approval processes.

How many lives have been saved by virtue of speedy drug approval since 1992? How many lives have been lost due to problems with drug safety that would have been caught in a slower and less rushed FDA approval process? These empirical questions are difficult to answer because

¹VERY VERY Drafty Version .160. Please not cite or circulate without permission. Jake Bowers is Assistant Professor, Department of Political Science and Center for Political Studies, University of Michigan and a Robert Wood Johnson Scholar in Health Policy Research at Harvard University (jwbowers@umich.edu). Ben Hansen is Assistant Professor, Statistics Department, University of Michigan (ben.b.hansen@umich.edu). Dan Carpenter is Professor, Department of Government, Harvard University. This project is a part of a larger project on the FDA which includes Chris Afendulis, Justin Grimmer, Susan Moffitt, Clayton Nall, and Evan Zucker. The authors appreciate the support and insightful comments provided to them by participants at the 2007 Northeastern Political Methodology Conference, at the workshops run by the Robert Wood Johnson Scholars in Health Policy Research Program at Harvard in 2006 and 2007 as well as at the national meeting of this group in 2006.

²The 1997 renewal was originally called the FDA Modernization Act (FDAMA), but as PDUFA was re-authorized again in 2002, and looks to be re-authorized yet again this year, the nomenclature has standardized on PDUFA I (in 1992), PDUFA II (in 1997), PDUFA III (in 2002) and now, pending, PDUFA IV (in 2007).

³See Carpenter et al. (2007) for evidence that (1) time to approval did speed up and (2) that approvals tended to “pile up” around the deadlines

they require comparison with alternate worlds; worlds to which we ordinary data analysts have no access. The first question in particular is so difficult that we do not address it in this paper. The second question, about the effect of the change in law on drug safety is more amenable to assessment without re-running history.

In fact at least two groups of scholars have tried to answer the second question: Olson (2002, 2004) predicted Adverse Drug Reaction (ADR) reports as a linear function of approval time using mostly drugs submitted for approval before PDUFA (i.e. drugs, most of which were not required to be approved by any particular deadline, but some of which were submitted after PDUFA I and therefore required to be approved within 12 months). She found that faster approval leads to discernibly *more* adverse drug reaction reports. Grabowski and Wang (2006) used a different sample of drugs, consisting mostly (but not entirely) of drugs submitted for review after PDUFA I, to fit essentially the same kind of linear model to ADRs and they found *no discernible relationship* between faster approval and ADRs.

Of course, the question of most interest to policy makers right now is not precisely whether approval time itself matters, but whether forcing regulators to work under deadlines matters for drug safety. In this paper we will be looking at the change in regulatory regimes rather than approval times, and will be training attention specifically on when particular drugs were submitted as an indicator of exposure to particular regulatory regimes. Thus, recent drug safety problems in the news, the current re-authorization of the law, and conflict between previous attempts to address the relationship between the new laws and drug safety, all motivate and justify our interest in this topic. In addition, the methodological conclusions we arrive at here will also be useful more broadly to the study of regulators and deadlines more generally, as well as for the analysis of other research designs where a discontinuity is available but where the samples are small.⁴

One caveat: If we find that safety was compromised, this does not mean that the current regulatory regime is bad. It is possible that more people were saved than killed. And it is also possible to use different normative frameworks to judge the overall success of such a policy. The point of this project is not to argue in favor for any particular normative calculus, but it is to lay to rest a particular empirical debate that has been going on about the effects of this policy. If we find that the new regime causes more safety problems than we ought to have expected from chance along, then others can use this fact to argue for or against the policy as they see fit.

In the rest of this paper we first discuss the methodological challenges that the FDA data pose and describe the data and research design that we have chosen to use. Then we estimate treatment effects, comparing approaches using different types of “pre-processing” (such as matching and discontinuities) to linear modeling practices and using two different frequentist modes of probabilistic inference. In the end, we will be able to discuss the merits of these different approaches to causal inference for this particular question, and will speculate a bit about their merits in more general cases.

⁴For formal theory on this topic see Carpenter and Grimmer (2007); Carpenter and Ting (2007). [Add page numbers about regression discontinuity designs from Shadish, Cook and Campbell (2001); Campbell and Stanley (1966) and other cites about this.]

2 THE DATA

2.1 Outcomes Relating to the Safety of Drugs

Outcome 1: Withdrawals from the world market due to safety problems. The current outcome of interest is whether a given new molecular entity (i.e. a new drug, not just an old drug in new packaging or an old drug put to new uses) is withdrawn from the US, Canadian or other markets for safety reasons.⁵ This outcome is much more rare than the Adverse Drug Reaction data studied by Grabowski and Wang (2006); Olson (2002, 2004): for example there were 13 such withdrawals among the 341 drugs for which we have submission date information submitted after PDUFA 1 was implemented (on October 1, 1992) up until the end of our data collection period (December 23, 2004) and 45 such withdrawals among the 1389 drugs for which we have valid submission date information submitted between February 20, 1943 and the implementation of PDUFA 1.⁶ Most of our analysis will focus on drugs submitted after 1962, when a new law was passed which required that no drugs be marketed without prior approval of the FDA (among other expansions of the FDA mandate).⁷

When a drug company pulls a drug off the market for safety problems it is (mostly) a clear signal that the drug is unsafe. However, some drugs which were known to be unsafe (such as Rezulin, a diabetes treatment, which killed people by poisoning their livers), stayed on the market because there were no safer alternatives. That is, being withdrawn for safety reasons means two things (1) the drug is unsafe, and (2) risks associated with this drug are too high given current alternatives and norms. Thus, the FDA approves many drugs which are dangerous but which are better than alternatives every year. That is, being withdrawn because of safety problems is a potentially weak (and rare) but clear measure of drug safety problems.

Outcome 2: The Adverse Drug Reaction data consists of millions of events mostly voluntarily reported to drug manufacturers by medical workers, and then reported to the FDA by the manufacturers under FDA requirements. Although there is a lot more variation in this outcome, finding an appropriate denominator which would allow comparisons across the drugs in terms of their safety has been difficult and fraught with data problems. We will add this analysis to the paper once we've surmounted the data challenges.⁸ However, even when we do that, our outcome will be quite noisy and with its own set of flaws: Are all such events reported to manufacturers by doctors? Probably not. How are the ones which are reported different from the ones which are not reported? Probably more severe, probably from more conscientious doctors, perhaps more from regions or hospitals where norms for this reporting are well established.

⁵We'll say "global market" for this outcome for the rest of the paper. But, some of these withdrawals only happened in the USA, or Canada, or only elsewhere.

⁶Our data set includes 2454 new drug applications which were eventually approved by the FDA and we know the submission date for 1730 (or about 70%) of them. We suspect, but are not absolutely sure, that these missing submission dates are all for drugs submitted before 1962, when the FDA gained its mandate to ensure efficacy and to approve drugs *before* they were marketed. We are missing approval date for all drugs for which we are missing submission date.

⁷Another caveat about our data: The data consist only of drugs approved by the FDA. The status of drugs submitted to the FDA which were not approved immediately, or never approved, is secret. We are not sure if even filing Freedom of Information Act requests for each such drug (even if we knew about them) would reveal this information to us.

⁸Thanks mostly to the insight of Chris Afendulis and the data of the Medical Care Panel Survey.

2.2 The Policy Change/Intervention/Treatment

The intervention that we care to assess is the change in policy, implemented as of September 1, 1992, which required that drugs be approved by a certain number of months after submission. Between September 1, 1992 and August 30, 1997, the deadlines were: 12 months for 90% of “standard” drug submissions and 6 months for 90% of “priority” drug submissions. After September 1, 1997 the deadlines for standard drugs tightened to 10 months. “Priority” drugs are drugs which members of Congress have deemed to be especially important or novel to receive fast approval. And “Standard” drugs are defined as “not priority drugs”. Clearly whether a drug is “standard” or “priority” will affect the impact of these changes in regulatory regime, so we will stratify by this status difference throughout the analysis.

We will assess the effects of PDUFA I (Sep 1, 1992-Aug 30,1997) and PDUFA II & III (Sep 1, 1997-present) separately. And we will focus only on PDUFA I for this version of the paper. That is the “treatment group” is, in theory, those new drugs submitted to the FDA for review between Sep 1, 1992 and Aug 30,1997.

Also, for the sake of this paper, we are not directly assessing the effects of being approved in a rush close to a deadline or being approved quickly close to the date of submission (with, we’d expect, less haste). That is, we are not looking directly at the effects of the “dose” of deadlines or haste received by a particular drug. Instead the analysis here focuses on the effects of the policy change (i.e. an Intent to Treat (ITT) analysis). If our ITT results are believable, and especially if our argument in favor of weak ignorability is strong, then we will be in a good position to continue the project by looking at speed of approval or approval close to deadlines using the policy change discontinuity as an instrument. This decision also means that, should we detect an effect of the policy change, we will not know the mechanism for the effect — it could be deadlines, but it could also be the user fees, the inexperience of new FDA reviewers, new technology systems at the FDA, or any number of other changes.

3 THE DESIGN

Here we formalize our desired inference within the potential outcomes framework (Neyman, 1990; Rubin, 1974, 2005) and clarify our assumptions.

3.1 Treatment Assignment (Z)

We have already defined the “treatment group” in 2.2. In this paper we will write $Z_i = 1$ to mean that a drug i was submitted to the FDA for review during PDUFA I and $Z_i = 0$ to mean that this drug was submitted to the FDA for review before PDUFA I was put into effect. We could (and will) also look at the effect of PDUFA II, in comparison to two different control groups (1) PDUFA I drugs and (2) pre-PDUFA drugs. But, we’ll start with just PDUFA I versus pre-PDUFA I.

“Assignment” is not quite the right word for the process by which drugs are submitted in time. However, we use words like “treatment”, “control” (which also isn’t right, since there is no control over treatment in this study), and “assignment” because (1) they clarify conceptual roles played by certain pieces of the research design and data analysis and (2) because we couldn’t think of better terms that would communicate as clearly these pieces of the design to people who are not

specialists in the FDA but who are specialists in applied statistics and methodology.⁹

3.2 Potential Outcomes (y_{Ti}, y_{Ci})

We want to know how many of the 9 drugs withdrawn from any global market for safety reasons which were submitted to the FDA under the PDUFA regime would have been withdrawn even if they had not been reviewed the deadlines (among other conditions associated with PDUFA I). Write $y_{Ci} = 1$ if a drug i would have been withdrawn if it had been submitted for review before PDUFA and $y_{Ci} = 0$ if it would not have been withdrawn, and write $y_{Ti} = 1$ to denote the case where that same drug would have been withdrawn if it had been submitted for review during the first PDUFA regime, and $y_{Ti} = 0$ if that drug would not have had safety problems leading to withdrawal if it had been reviewed under the first PDUFA regime. Thus we have one pair of “potential outcomes”, y_{Ti}, y_{Ci} for each drug in the dataset — which consists of all of the drugs submitted for approval to the FDA since about 1948.

3.3 Treatment effects ($\tau_i = y_{Ti} - y_{Ci}$)

For a given drug, we define our treatment effect in terms of a counterfactual statement: $\tau_i = y_{Ti} - y_{Ci}$, which is the difference in potential outcomes, the difference in potential for a drug to be withdrawn from the global market because of safety problems. If we can require $y_{Ti} \geq y_{Ci}$, the estimand that we desire across all of the drugs, call it A , is $A = \sum_i Z_i \tau_i = \sum_i Z_i y_{Ti} - y_{Ci}$: which is the number of drugs which, reviewed under PDUFA, would have been withdrawn for safety reasons and which reviewed prior to PDUFA would have not have been so withdrawn. We can require that $y_{Ti} \geq y_{Ci}$ because we imagine that PDUFA only increased the probability of such safety related problems.¹⁰ For now, this assumption is substantively reasonable: no one is claiming that PDUFA increased safety, only that it hasn’t harmed safety (or even that if it did harm safety, more people are being saved from speedy approval than killed from drug safety problems).

3.4 Varying Treatment Effects: Attributable Effects

Notice that in the preceding definition of a treatment effect we wrote τ_i and not τ . It is common to define the treatment effect as something which is additive and constant such that, say, $y_{Ti} = y_{Ci} + \tau$. But with binary responses ($y \in \{0, 1\}$) the constant, additive treatment effect assumption is not tenable. For example, assume the additive, constant treatment effects model. Then state that $\tau = .1$, what then is implied about y_{Ti} and y_{Ci} ? If $\tau = .1$ or $\tau = 2$ or τ is anything except for $-1, 0$ or 1 , then the potential responses are forced to take on values other than 0 or 1 . Even if $\tau \in \{-1, 0, 1\}$, each such value creates rigid requirements for y_{Ci} that is, $\tau = 1$ requires that *all* drugs have the *same* potential response to control, $y_{Ci} = 0$.¹¹ However, if we use our conceptual definition of treatment effects as additive but varying, $y_{Ti} = y_{Ci} + \tau_i$, then we face no such conceptual difficulties.

⁹Rosenbaum(xx) suggests “referent” rather than “control.” But this usage is not common and we worry it would confuse people.

¹⁰This assumption is called “nonnegativity” by Rosenbaum in [cite]. If this is not the case, we will know it by virtue of an empty confidence set (foreshadowing our estimation method), and, in that case we can easily then estimate treatment effects under a “nonpositivity” assumption as well. Since, tests of hypothesis that require $y_{Ti} \geq y_{Ci}$ and test of hypothesis that require $y_{Ti} \leq y_{Ci}$ are mutually exclusive allowing additional testing of this kind of does not artificially increase the probability of Type I errors.

¹¹See Rosenbaum (2002b, Chapter 5) for more discussion of why a binary response is incompatible with a constant, additive treatment effects assumption.

Our estimand of interest is $A = \sum_i Z_i \tau_i$, which is the sum of the drug-specific treatment effects which themselves vary across the sample of drugs.¹²

3.5 Ignorability

Although our estimand emerges nicely from the potential outcomes framework in the context of a binary response, the preceding discussion assumed that merely summing individual treatment effects would provide an unbiased estimate — an estimate which only reflected the comparisons of potential responses, not other comparisons. That is, we assumed that we could ignore other possible confounds or covariates, $\mathbf{X}$, when we made our comparison of potential outcomes. This assumption, known as the “strong ignorability” assumption and can be formalized thus (following (Rosenbaum and Rubin, 1983a) but using the notation of this current paper):

$$y_T, y_C \perp\!\!\!\perp Z | \mathbf{X}, 0 < p(Z = 1 | \mathbf{X}) < 1 \quad (1)$$

That is, the potential outcomes are independent of treatment assignment once covariates have been taken into account. In the current data example, this assumption would imply that drugs of type A which have a high prior probability of having safety problems are no more likely to be assigned to treatment (i.e. reviewed under PDUFA) than drugs of type A which have a low prior probability of having safety problems. This is hard to imagine in this research design. Drugs from the 1950s probably have different potential to have been withdrawn from the global market because of safety reasons than do drugs from the 1990s.

As with most studies where the assignment mechanism is not controlled, we cannot make assumptions like this easily in this study. However, also like most other observational studies, our task in estimating a causal effect involves *choosing observations* and *organizing our data* such that we can make as strong a case as possible in favor of this assumption. The general strategy for making this case is to (1) choose a sample of drugs to compare which are, as far as we can observe, as similar as possible in terms of characteristics of the drugs which are not affected by regulatory regime (i.e. pre-treatment covariates that matter for potential outcomes); and (2) given this choice of sample, organize the sample into sets which are homogeneous in terms of those characteristics. We aim to make the final organization of our data look as much like a randomized experiment as possible.

Why try to make an observational study look like a randomized, controlled, experiment? We want to do this because one important purpose of randomized assignment of treatment (as is done in experiments) is to allow analysts to claim that potential outcomes are independent of treatment assignment. Of course, ignorability does not hold without question in every instance of an experiment, but, in large samples and across multiple realizations of an experiment, the act of random assignment ought to ensure ignorability. Thus, one way to make the argument that our observational study data have been chosen and organized such that we don’t have to worry that comparisons of outcomes are confounded by other comparisons, is to compare our study to a randomized experiment. If, in the end, we cannot reject the null hypothesis that our data were generated by an ignorable assignment process like that used in a lab experiment, then we will have gone far enough to estimate a treatment effect under that assumption. Of course, since our data were manifestly *not* generated by a random assignment process (at least across all 2500

¹² (See Rosenbaum, 2002a, 2001, for more details and development of this estimand) and for an application in political science see (Bowers and Hansen, 2006a)

drugs), we will want to conduct a sensitivity analysis to gauge the impact of factors which might be confounding our comparison; factors which we cannot and do not observe in our data.

3.6 Choice: Using Discontinuities to Make the Case for Ignorability

For the purposes of this study we can use the discontinuities in regulatory rules which occurred on September 1, 1992 and September 1, 1997 in order to help bolster the case for ignorability.¹³ We have already suggested that comparing potential outcomes from a drug from the 1950s to a drug from the 1990s would probably involve many confounds. However, we imagine that a drug submitted the day before the new rules were enacted would not be that different in terms of safety problems a priori from a drug submitted the day after the new rules were enacted. If this were the case, then we would (1) know that the two drugs received different interventions (Z) and (2) that the probability of receiving the treatment (i.e. being submitted before or after PDUFA) would not depend on safety based differences between the drugs. If that was so, then we could substitute y_{Cj} for the pre-PDUFA drug j for y_{Ci} in our estimate of the effect of treatment on unit i , so now we would have $\hat{\tau}_i = y_{Ti} - y_{Cj}$ knowing that $y_{Ti} - y_{Cj}$ captured only differences in potential outcomes and nothing else.

How might we make ourselves (and our readers) comfortable with the idea that we can choose drugs submitted *close* to the change in the law and thus have sets of drugs for which comparisons would be less confounded than they would be if we hadn't so chosen comparison sets? First, by knowing something about drug trials and the passage of this law. And second, by looking at the data to see if the drugs submitted just before the rule change tended to be different (in terms features which we imagine would predict safety, like the disease the drug was meant to treat, or the number of safety related mentions in academic/medical articles in Medline, etc...). That we, we will present some *prima facie* evidence and also test for balance.

3.6.1 Prima Facie Evidence in Favor of Ignorability

Is there any reason to believe that drugs submitted right before PDUFA I was implemented are different in terms of safety than drug submitted right after PDUFA I? Not really. The law itself was passed in September of 1992. There were no mentions of a proposed law in the news media until about 3 months before it was passed. Knowledgeable legislative observers at the time noted that the pharmaceutical industry “came around” to the idea of user fees in exchange for speedy review and approval only at the last minute. Knowing this about the process by which the law was passed, add to this knowledge that Phase III clinical trials (the clinical trials that immediately precede submission of a new drug application to the FDA) typically take between 9 and 18 months, with a methodology and schedule agreed upon with the FDA in advance. All firms desire to rush through their trials as fast as possible, regardless of the regulatory environment, because of competition and opportunity costs and the expiration of patents (patents are often granted *before* clinical trials begin, and so are often nearly expired by the time that clinical trials

¹³Notice that we are not talking about whether the rules were actually or uniformly applied here. We have some sense from testimony of FDA officials in the Federal Register that the PDUFA rules were “phased in” somehow although we don't yet know the decision criteria used to assign PDUFA rules to which drugs when they were submitted. We also learned recently that there were informal attempts by the FDA to “deal with the backlog” of cases for which approval was pending as of the implementation of PDUFA I. So, again, until we get more detailed information about which drugs were subject to particular speed and which were not, we can only treat these data as an “encouragement design.” That is, we are doing an “intent to treat” analysis here we don't know about “compliance” with this treatment yet.

are over). An interview with a current pharmaceutical company FDA liaison, who was previously an FDA reviewer, confirms this fact: firms were either too tied down by the schedule and intricacies of Phase III trials or were too busy racing patent clocks and competitors to market to be able to delay or speed submission of a new drug application merely because of the change in regulation.

3.6.2 A Data Based Evidence in Favor of Ignorability

What do the data say? Is it the case that if we choose sets of drugs submitted “close to” the date of PDUFA implementation, that these drugs somehow look similar? We used three measures to summarize similarity between drugs submitted during the PDUFA I regime and drugs submitted to the FDA before that regime was implemented. And we applied these three measures to 6 different definitions of a temporal window around the PDUFA I implementation date: 5 windows which increased symmetrically in size from 1 year to 5 years, and a 6th window which included drugs from 1962 up until PDUFA I implementation as “controls” and drugs submitted during all of the PDUFA I regime (from September 1, 1992 until August 30, 1997) as “treatment”. If our hunch about the discontinuity is right, we’d expect to see that the differences between treated and control groups should diminish the closer to the discontinuity the windows are drawn.

Our first measure of distance is a standardized difference of means adjusted for the “standard” versus “priority” status of a drug following the development in Hansen (2006a, §3.2). For a binary covariate, one can think of this measure as the difference in proportions of members of the treatment group having $x = 1$ and members of the control group having $x = 1$ weighted by the size of the strata (where, strata are “standard” drugs versus “priority” drugs in this case).¹⁴ For more general measures the differences are a similarly weighted difference of means. Second and third, we used order statistics describing the difference in the empirical cumulative distributions of a given covariate; specifically the median and maximum difference between the distributions for the control groups and the treated groups.¹⁵

These different distance metrics were calculated separately for approximately 400 covariates including: Media Coverage Before PDUFA I, Incidence of Primary Indication, Indicator for Incidence of Primary Indication Missing, Death Rate for Primary Indication, Indicator for Death Rate for Primary Indication Missing, # Hospitalizations for Primary Indication, Order of Entry Into the Market, # Submissions by Firm, Indicator for # Submissions by Firm Missing, and two measures counting the number of mentions of the drug during clinical trials in Medline: one capturing overall mentions and the other counting mentions alluding to the safety of the drug. In addition, the covariates included indicators for the firm of submission and the disease the drug was meant to treat (aka the “primary indication”). Rather than show plots or tables with hundreds of columns, Figure 1 plots the distributions of those scores across the covariates. For this draft of the paper, we recommend viewing this figure in color and we invite ideas for how to make it easier to see in grayscale printing.

The leftmost panel of Figure 1 shows the distributions of the standardized mean differences, adjusted for standard versus priority status. This figure does not show appreciably tighter clustering around zero as the window tightens — for example, the smallest window (at 1 year, drawn

¹⁴The specifics of the weights are actually $\bar{m}_b / k_b$ where, $\bar{m}_b$ is the average size of the b^{th} strata and k_b is the total size of the sample (See Hansen, 2006a, §5, for an argument about why these weights are optimal in a certain sense)

¹⁵This following the discussion in Sekhon (2006) and Imai, King and Stuart (2006).

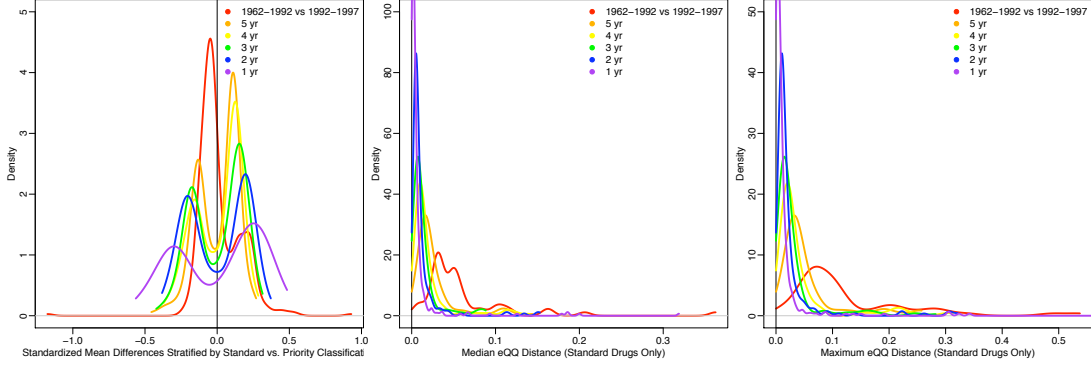


Figure 1: Distributions of pre- vs. post-PDUFA differences using three different metrics. Densities calculated using the `density` function in R with all arguments at default values except that we adjusted the bandwidth to be twice as smooth as the default and limited at the left and right to the observed values of the data.

in purple) produces a wider and flatter distribution than any of the 6 windows. Of course, allowing 30 years of drug submissions into the control group produces a very wide variance (the red distribution), but also fairly closely peaked around zero. In the end, this result is not surprising: there ought to be both less bias near the discontinuity, but more noise as well since the sample of drugs becomes quite small (with 26 drugs submitted in the year before PDUFA I and 23 drugs submitted in the year after PDUFA I; but 98 pre- and 128 post- for a 4 year window and 935 pre- and 163 post-PDUFA for the largest group allowing comparisons back to 1962).

The middle and the right panels of Figure 1 summarize the difference in empirical quantiles between the treated and control group, with the middle panel showing the median difference and the rightmost panel showing the maximum difference; statistics that ought to be less sensitive to the sample size. In each of these cases, the differences collapse toward zero as the sample window tightens — although there is appreciable variation in the tails of these densities.

An omnibus balance assessment Another way to put this question is to ask whether we can distinguish the differences among drugs from those that we could expect had the submission date of drugs been randomly assigned (in this case, within strata of “standard” versus “priority” drugs). Since the distribution of the standardized, adjusted, mean differences is well defined under this null hypothesis, we can easily compare our observed differences to that null in order to assess and summarize the evidence in favor of this particular imagined state of the world. One can develop a test for the linear combination of all covariates tested. Such an omnibus test developed in Hansen (2006a, §4) and implemented in the `RIttools` software package (Bowers and Hansen, 2006c), is what we used to gauge our observed set of differences against what we know the world would look like if these covariates had in fact been randomly assigned to be reviewed before versus after the implementation of PDUFA I. If we what we observe is consistent with the fantasy that our data were randomly assigned, then this is good evidence in favor of the treated and the control observations being very similar; perhaps similar enough that we would feel comfortable claiming that we believe in the ignorability assumption that is required for the calculation of our treatment effect. (Or perhaps, just similar enough that we can feel more confident with our attempts to use other covariate information (other than year of submission) into order to organize our data into sets that even more closely approximate what we would like to see if we had run this study as an experiment.)

To this end we started with the treatment group defined as all those new drugs submitted during PDUFA I. Then we inched out the control group window for each unique date in the dataset — i.e. we checked for balance on those roughly 400 covariates (or those which had non-missing information for a given definition of the control group). For each such definition of our control group, we recorded the p-value for the omnibus test of balance alluded to above. Figure 2 displays the results of this exercise. The thin horizontal lines encode some p values which are commonly understood provide thresholds, beyond which people often assume the evidence is incompatible with the null. As we can see, once the control group includes drugs submitted earlier than about 1988, the test ceases to provide much evidence in favor of our randomization illusions.

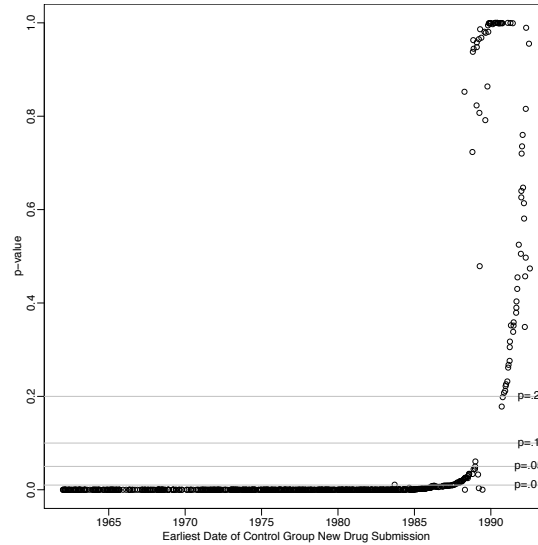


Figure 2: P-values from an omnibus χ^2 of the null hypothesis that the linear combination of all listed covariates was assigned at random to drugs submitted before PDUFA I but after the date specified on the x-axis versus drugs submitted during PDUFA I (See Hansen (2006a, §4) for development of this test.)

3.6.3 What do we make of the evidence for balance based on design?

The prima facie evidence, the graphical displays of distance metrics, and some probability based summaries of distances suggest that small samples of drugs submitted just before and after the discontinuity induced by the change in law do look roughly similar (for example, very few mean differences more than ± 1 sd). We have some sense that such differences are more or less what we might have expected due to chance. However, these analyses make clear that merely choosing a 4 or 5 year window is not enough to make the argument for ignorability as strong as we'd like it to be. And choosing a 1 year window may provide us with too little data to analyze.¹⁶ Luckily, given the number of covariates that we have available, we can affirmatively adjust and organize the data above and beyond our choice of sample. We hope that by doing so we will make an even stronger argument that our eventual treatment effect estimate reflects differences in potential outcomes and, as much as possible, not other kinds of confounding differences.

¹⁶For example, only 1 safety based withdrawal occurred among the drugs submitted 1 year after PDUFA I — and only 1 safety based withdrawal among the drugs submitted one year before PDUFA I.

4 MATCHING BASED ADJUSTMENT: USING PROPENSITY SCORES, PROGNO- SIS SCORES, AND FULL MATCHING TO FURTHER BOLSTER THE CASE FOR IGNORABILITY

First, a brief aside:

4.1 Bias and Variance for $\hat{\beta}_{OLS}$

Say, $\mathbf{y} = \beta\mathbf{X} + \mathbf{e}$ where

$$\text{var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \quad (2)$$

$$\mathbb{E}(\hat{\beta}) = \beta + \mathbb{E}(\mathbf{X}^T\mathbf{X})^{-1} \cdot \mathbb{E}(\mathbf{X}^T\mathbf{e}) \quad (3)$$

Remember that (2) is a function of σ^2 (i.e. how well the regression line fits the $\mathbf{y}$ data; a summary of the tightness or spread of y_i around the fitted regression line) and $(\mathbf{X}^T\mathbf{X})^{-1}$ (i.e. the variation in each individual covariate as well as the linear relationships among the covariates). Thus, more precise estimates of β , are based on less residual variation in the outcome (i.e. smaller σ^2) and more covariates, with more variation (and less covariance) (i.e. “bigger” $\mathbf{X}^T\mathbf{X}$).

Equation (3) reminds us that the bias in $\hat{\beta}$ also depends on the variances of and covariances among the covariates as well as the covariance between the covariates and $\mathbf{e} = \mathbf{y} - \beta\mathbf{X}$ (i.e. the pieces of $\mathbf{y}$ that are not a linear function of $\mathbf{X}$). So, in order to have confidence that our estimate of β is not misleading, we need to make $\mathbb{E}(\mathbf{X}^T\mathbf{e})$ as small as possible¹⁷

Why talk about the small sample properties of OLS as a smoother (not as causal estimand) in a paper on matching? We do this because the general principles are the same, whether we estimate treatment effects using $\hat{\beta}$ or some other way — in all cases we desire tight confidence intervals; intervals which depend on both the variation in the outcome which is not somehow accounted for by either our modeling assumptions or by design based adjustments as well as the variation/covariation induced by these decisions (i.e. such as the design to include another linear “control” variable, or to stratify in a particular way). And we desire to avoid confounding as much as possible (i.e. to reduce something akin to $\mathbb{E}(\mathbf{X}^T\mathbf{e})$ to zero).

In our particular case, as we have intimated so far, we have a small amount of data. And so, we would like to produce estimates that are as precise as possible (of course, who doesn’t desire this? It is just that in our case, if we are not careful we may not have the power to detect any effects at all — which may mean that we should just give up on this enterprise; or which may mean that we should think just a bit harder about how to increase our precision without hurting our bias reduction efforts). Most of the strategy in causal inference inspired by Rubin, Rosenbaum, and their collaborators has focused exclusively on bias, and has also mainly been concerned with adjustments that balance covariates with regard to assignment. However, as this little aside reminds us, balance with regard to the response also matters. Of course, one does not have as free a hand when seeking to balance covariates with respect to response as with assignment [more]. In what follows we attempt to pay attention to the response, but to do so in such a way that is consistent with the general principles of causal inference workflow.

¹⁷We’d also like $\mathbb{E}(\mathbf{X}^T\mathbf{X})$ to be as large as possible. If this occurs without stopping as the sample size increases, then we’ll have a consistent estimate even if we have a biased one. See Achen (1982) for a nice proof of this result.

4.2 Back to the Paper: Prognosis Scores

Hansen (2006b, 2007) proposed to produce a score predicting y_C , the potential response to control, using only information from controls. Here we will try to offer some intuitions about the score. More details and theory are available in Hansen (2006b).

Remember that a propensity score ($\phi(\mathbf{x}) = E(Z|\mathbf{X} = \mathbf{x})$), correctly specified, allows a scalar $\phi(\mathbf{x})$ to replace the whole matrix of potentially confounding covariates $\mathbf{X}$ during adjustment with matching or in linear models (Rosenbaum and Rubin, 1983b, 1984). That is, $\phi(\mathbf{x})$ captures all of the information in $\mathbf{X}$ that can confound comparisons of y_T and y_C based on Z . This is a very convenient (theoretical) attribute of propensity scores — that you can replace many covariates with a single measure. Another way to put this property would be to say that $\phi(\mathbf{x})$ is a propensity score if:

$$\mathbf{X} \perp Z | \phi(\mathbf{X}).$$

That is, if you have specified the propensity score correctly (i.e. included all of the covariates, and with the right functional form to predict treatment assignment), and if you match subjects with the same propensity score to each other, within those sets you can ignore $\mathbf{X}$ when you compare y_C to y_T .¹⁸

A prognostic score is similar to a propensity score in that it is also a function of covariates, $\mathbf{X}$. But it produces balance on outcomes among the control group, such that:

$$y_C \perp \mathbf{x} | \psi(\mathbf{X})$$

Notice that we are only modeling the outcome among the controls. In the case of our study, we go further than this — only modeling the outcome among drugs submitted to the FDA for review before our “control” group was defined (i.e. before 1988, but not between 1988 and 1992 which are the years defining our control group.)

[more]

4.2.1 Why use prognostic scores?

Prognostic scores help us with this analysis in two ways. First, they help us overcome the problem of separation on propensity scores. When we fit a propensity model using a logistic regression we found that the model produced scores where the scores for the treated units were quite different from the scores for the control units. This led to problems with organizing the data into subsets which were homogeneous on propensity scores — even full matching using a form of logistic regression allowing for shrinkage across the coefficients excluded 5 drugs (using a wide caliper of 3 sd). That is, what we really want within matched sets is for potential outcomes to be equivalent. However, by using the propensity score, we create sets where covariates that matter for treatment assignment are equivalent — and these covariates, we know, cannot then influence the comparison of observed outcomes. However, some of these covariates (or many of them in our case) have no relationship with the outcome, although they may add something

¹⁸Notice that so far I have not referred to actual observed responses R_i , which are realizations of one of the potential outcomes, but not both, such that: $R_i = Z_i y_{Ti} + (1 - Z_i) y_{Ci}$. What this means is that among members of the control group $Z_i = 0$, and so $R_i = y_{Ci}$ and among members of the treatment group $R_i = y_{Ti}$.

to predicting treatment assignment. Including the prognostic scores into the propensity model allowed us to produce more matched sets without throwing away information, even using the same caliper. That is, it solved (in part) our problem with separation.

Second, prognosis scores ought to help reduce the residual variation in the response (i.e. do the equivalent of making σ^2 smaller). This ought to increase our power to detect effects. Power is especially important in this particular case since we have so few positive outcomes and our samples sizes are in general quite small.

4.2.2 How did we use prognostic scores?

First, we calculated prognostic models of y_C . We estimated three models of the observed response to the control condition $R_i|Z = 1$ only among drugs submitted before our sample window was defined. The models are allowed for different coefficients for priority and standard drugs for the following covariates: Media Coverage Before PDUFA I, Incidence of Primary Indication, Indicator for Incidence of Primary Indication Missing, Death Rate for Primary Indication, Indicator for Death Rate for Primary Indication Missing, # Hospitalizations for Primary Indication, Order of Entry Into the Market, # Submissions by Firm, Indicator for # Submissions by Firm Missing, and two measures counting the number of mentions of the drug during clinical trials in Medline: one capturing overall mentions and the other counting mentions alluding to the safety of the drug. The drug and firm indicators were not included here because, several of the drugs and firms extant in the pre-1988 data did not exist in the analysis sample (1988 to 1996). The first model was a simple linear interactive model with 28 terms. The second model elaborated on the first model by allowing 6 piece natural, cubic, regression splines for all of the continuous variables in the first model — yielding 100 terms. And the third model added complexity to the first model by allowing all two way interaction terms — leaving 106 terms. Each of these models amounts to a different projection of $\mathbf{X}$ onto the space of y_C . Or, a different dimension reduction. Thus, each one has a slightly different bit of information which can lead to a further reduction in separation and further increase in power.¹⁹ In the end, one important point of using prognostic scores is to “wring” as much information out of $\mathbf{X}$ as possible — not only the information relating $\mathbf{X}$ to $\mathbf{Z}$, but also the information related to y_C .

Once we had the three different vectors of coefficients representing three different attempts at reducing the dimension of $\mathbf{X}$ as related to response, we applied those coefficients to the analysis sample data. That is, we merely took those coefficients and predicted y_C using those coefficients. We did not re-estimate the models on the analysis data.

[show the models]

¹⁹In future work we hope to have a more principled way to generate different prognostic score models, and also to extract functional forms and coefficients from them. For example, in this paper we just use the MLEs, but we could also take different vectors of coefficients from the posteriors; vectors which summarize the posteriors in different ways.

Second, we estimated propensity models of Z conditional on our prognostic scores. Then, we estimated a propensity score model using those predicted y_C scores as well as dummy variables for drugs, and some of the variables that seemed especially relevant to whether the drug might have safety problems: Incidence of Primary Indication, Indicator for Incidence of Primary Indication Missing, Death Rate for Primary Indication, Indicator for Death Rate for Primary Indication Missing, # Hospitalizations for Primary Indication. These propensity scores exhibited less separation than the propensity scores which did not use the prognostic scores.²⁰

[show the models]

Figure 3 shows the smoothed distributions of the propensity score models with and without the prognostic scores, calculated with and without shrinkage. Each plot is drawn to include the entire range of the score. So, for example, the scores for the “plain old” propensity score (shown in the top left panel) range from around -3 to around 11 on the scale of the linear predictor ($\hat{\beta}X$) while that same model calculated to allow coefficients with low predictive power to be smaller (shown in the top right panel) had scores ranging from about -5 to about 5.

The plots using shrinkage in the estimation of the propensity score show configurations of scores which will make matching (or other adjustment) easier. That is, these scores mean that there will be few treated subjects who are very distant from any control subject (and vice-versa) on the metric of the score. And the combination of the prognostic scores within the propensity score models enhances this feature (as shown by the bottom right hand panel). The increased variance in the density of the score among the controls means that we can make more distinctions among the members of the control group; distinctions which will enable more flexibility when it comes to adjustment.

Third, we matched treated to control units based on the propensity scores. Finally, we matched drugs submitted in the four year window post-PDUFA I to the drugs submitted in the four year window pre-PDUFA I using matrices containing the absolute differences in propensity scores between drugs (stratified by standard vs. priority status). We chose to do full matching without replacement as implemented and described in Hansen and Klopfer (2005); Hansen (2004). Full matching recommended itself to this particular application because (1) we desired not to exclude any units within the 4 year windows a priori and (2) and because matching without replacement allows flexibility in using the matched sets in treatment effect estimation. In addition full matching is a global procedure: this means that we did not worry about the initial sorting of our data since the procedure looks for a set of matches which minimizes the total set of distances between treated and control units rather than looking set by set to minimize such distances.

Table 1 and 2 show the configuration of sets of two matchings: one (in Table 1) uses the propensity score shown in the top right of Figure 3; and the other (in Table 2) uses the propensity score shown in the bottom right of that same figure. The set involving prognostic scores has more sets (70 versus 57), excludes fewer observations (0 versus 1), and the sets are less un-even in size. Since, in theory, more sets, with more observations, with less variance in set size, are all characteristics which ought to lead to tighter confidence intervals.²¹

²⁰Estimated again, using the `br1r` package for R (Firth, 2006); a package which allowed for “bias-reduced” logistic regression — essentially a way to allow shrinkage to occur across coefficients such that coefficients which contribute very little to the forecast receive less weight.

²¹See Hansen (2004) for some discussion and examples of trying to get `fullmatch` to produce sets which are more even in size.

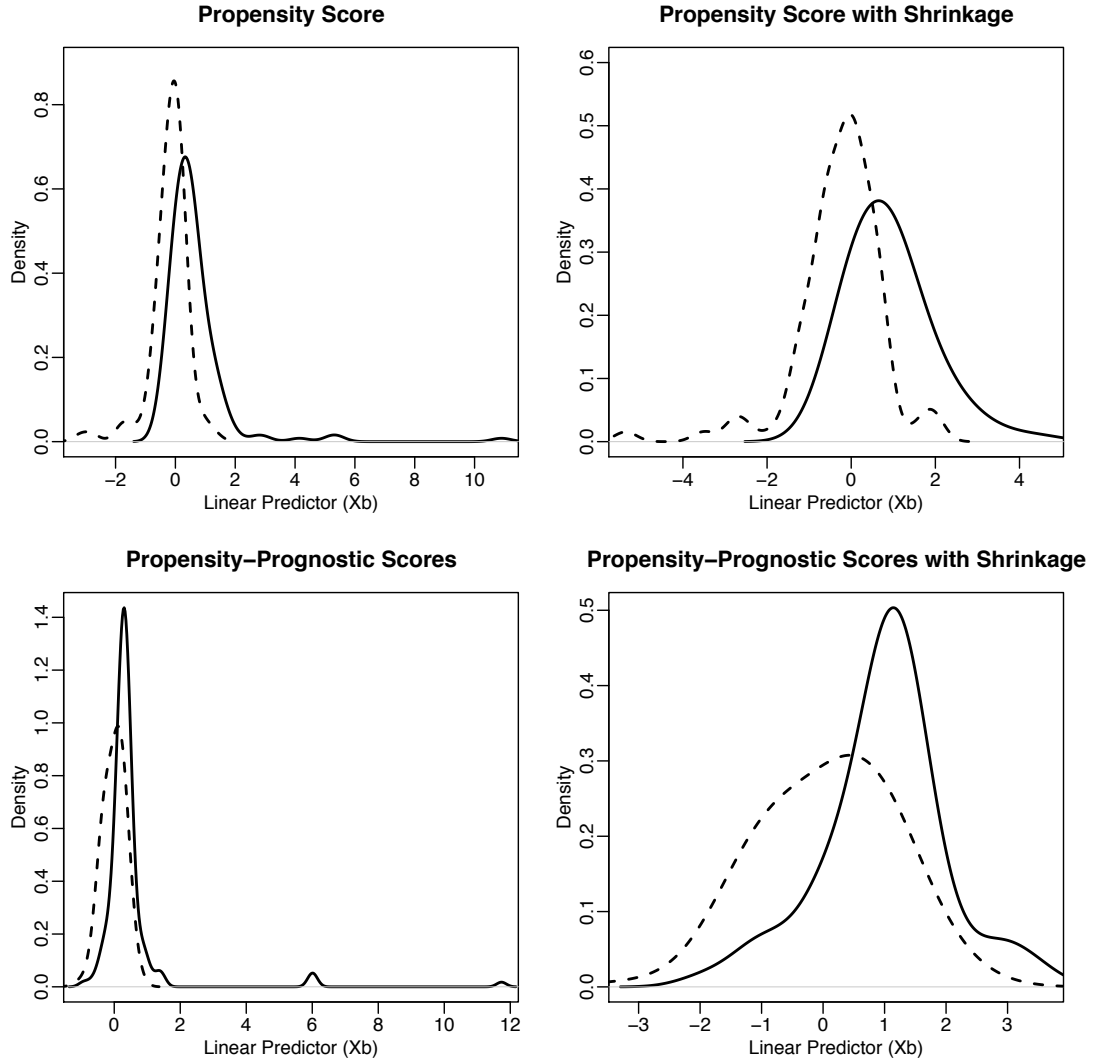


Figure 3: Distributions of 4 different kinds of propensity scores plotted for pre-PDUFA drugs (dotted lines) versus post-PDUFA drugs (solid lines) for drugs submitted within 4 years of PDUFA I. Propensity scores were calculated using logistic regression (`glm`). Shrinkage was imposed across the covariates by using a Jeffreys prior (implemented using penalized mle in `br1r` (Firth, 2006)). Densities calculated using the `density` function in R with all arguments at default values except that we adjusted the bandwidth to be twice as smooth as the default.

Priority			Standard		
Set	Pre	Post	Set	Pre	Post
St.01	1	0	Pr.1	2	1
St.1	1	4	Pr.10	1	14
St.10	1	1	Pr.11	1	1
St.11	1	3	Pr.13	1	2
St.12	1	7	Pr.15	1	5
St.13	1	6	Pr.18	1	2
St.14	1	1	Pr.2	2	1
St.2	1	19	Pr.23	1	8
St.20	1	13	Pr.3	1	2
St.21	1	1	Pr.30	7	1
St.24	1	3	Pr.31	2	1
St.25	2	1	Pr.32	1	1
St.28	1	3	Pr.33	2	1
St.29	2	1	Pr.34	2	1
St.32	1	1	Pr.39	1	1
St.33	1	2	Pr.4	1	1
St.46	1	1	Pr.40	1	1
St.48	1	1	Pr.41	1	1
St.5	1	16	Pr.42	1	1
St.50	1	1	Pr.43	2	1
St.51	2	1	Pr.5	1	1
St.53	1	1	Pr.6	2	1
St.59	1	1	Pr.7	1	1
St.61	4	1	Pr.8	1	1
St.68	1	1			
St.69	1	2			
St.75	1	1			
St.77	20	1			
St.8	1	5			
St.89	2	1			
St.90	2	1			
St.93	1	1			
St.95	2	1			

Table 1: Matched sets from a full matching using propensity scores with shrinkage. The matching relied on the `optmatch` package (Hansen and Klopfer, 2005). Calipers of 3 sd excluded 1 drug (seen above as “St.01”).

Priority			Standard		
Set	Pre	Post	Set	Pre	Post
St.1	1	15	Pr.1	1	1
St.10	1	1	Pr.10	1	2
St.11	1	1	Pr.11	1	1
St.12	1	2	Pr.12	1	7
St.13	1	4	Pr.13	1	1
St.14	6	1	Pr.15	1	1
St.16	1	1	Pr.16	1	1
St.17	1	3	Pr.17	1	2
St.18	1	1	Pr.18	1	1
St.19	1	1	Pr.2	1	6
St.20	1	3	Pr.25	1	1
St.21	1	2	Pr.28	1	2
St.25	1	1	Pr.29	1	1
St.26	1	1	Pr.3	1	1
St.29	1	1	Pr.32	1	3
St.30	1	4	Pr.33	1	1
St.31	1	4	Pr.34	1	1
St.35	1	4	Pr.36	1	4
St.4	1	13	Pr.38	1	2
St.45	1	5	Pr.39	1	2
St.46	1	1	Pr.4	1	2
St.48	1	1	Pr.42	1	1
St.49	1	1	Pr.43	1	1
St.50	1	2	Pr.48	1	1
St.53	1	2	Pr.5	1	2
St.55	1	1	Pr.6	1	1
St.56	1	1	Pr.7	6	1
St.63	1	1	Pr.8	5	1
St.64	1	1			
St.65	1	1			
St.68	1	1			
St.69	2	1			
St.78	1	1			
St.8	1	12			
St.85	1	1			
St.86	1	1			
St.87	1	1			
St.89	6	1			
St.90	1	1			
St.91	1	1			
St.96	7	1			
St.97	3	1			

Table 2: Matched sets from a full matching using propensity scores calculated using prognostic scores with shrinkage. Calipers of 3 sd did not exclude any drugs from the matching.

Finally, we checked the propensity balance of the matched sets. When we tried to decide how large of sample of drugs to use to estimate treatment effects, we compared the differences on covariates between pre- and post-PDUFA I drugs to the differences we would expect to see if someone had randomly assigned submission date (pre- vs. post-) to those sets of drugs. We take the same tack here. This time the adjustment is not only stratification by priority versus standard status, it is membership in matched set. Figure 4 shows, graphically, how the differences between covariates got smaller as a result of the matching (using the matching shown in Table 2). The omnibus test of balance described earlier provided some evidence against the null of random assignment on the unadjusted data ($p=.04$) but no evidence against this null when used on the data taking into account matched sets ($p=.9$).

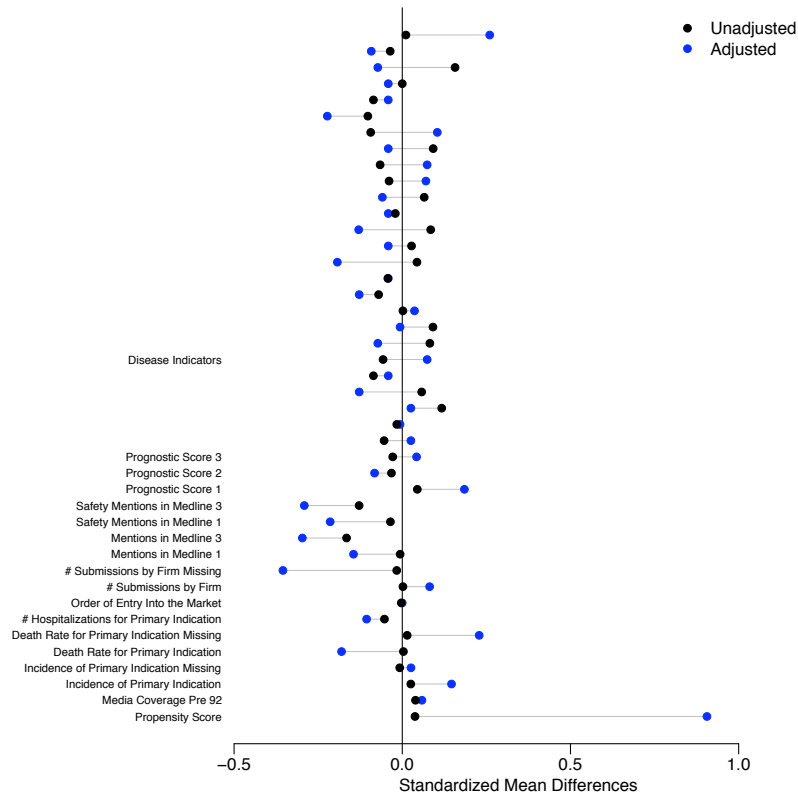


Figure 4: Standardized, mean differences unadjusted before matching (blue dots) compared to after adjusting via matching (black dots) on a list of covariates plus the propensity score including prognostic scores estimated allowing for shrinkage (the bottom dot).

Summary Although we began this paper talking about ignorability, now we are talking, some, at least, about separation. It turns out that separation is an assumption involved in the ignorability statement:

$$y_T, y_C \perp\!\!\!\perp Z|X, 0 < p(Z = 1|X) < 1 \quad (4)$$

The statement that $0 < p(Z = 1|X) < 1$ means that no individual covariate (and no set of covariates) perfectly predicts treatment assignment. In our case, the set of covariates did a very good job of predicting treatment assignment — such a good job that our estimates of $p(Z = 1|X)$ could get very very close to 1 or 0. Using the logistic regression allowing for shrinkage and also the prognostic scores allowed us to more fully believe that $0 < p(Z = 1|X) < 1$.

In addition, this combined score produced a configuration of matched sets which, in theory, ought to allow for tighter confidence intervals: more sets, less variation across sets, fewer excluded units.

4.2.3 Questions about This Procedure

Why is it ok to condition on $\hat{y}_C$? [TO DO]

Why embed the prognostic scores within a propensity score model? In past work we have tried using prognostic scores merely to stratify the data (without matching) (Bowers and Hansen, 2006a), to estimate heterogeneous treatment effects (Bowers and Hansen, 2006b), and also wrapped into a Mahalanobis distance along with the propensity score (having equal weight with the propensity score in that distance calculation) (Bowers and Hansen, 2005). All of these approaches, including embedding the prognostic scores inside a propensity score equation (as we’ve done here) are acceptable.

[explain more]

What about the uncertainty in the estimation of the prognostic scores? This question could also be asked about the propensity scores themselves; and it is an open question on both accounts. Our method of estimating treatment effects, however, implicitly accounts for both of these types of uncertainty — accepting a particular null hypothesis because there is not enough available disconfirming evidence could be interpreted to mean that (1) the treated group and the control group are quite close together in terms of their potential outcomes and/or (2) that we just have too much noise/too little data to distinguish this small difference from zero.

5 PERMUTATION BASED INFERENCE FOR THE ATTRIBUTABLE EFFECT, A

There were 9 safety based withdrawals in the 4 years after PDUFA I. And 5 in the 4 years before PDUFA I. Comparing 9 to 5 might lead one to conclude that PDUFA I had some influence on increasing safety problems. Of course, if we scale these numbers of withdrawals by the numbers of submissions (98 and 154), then we are comparing 5.1% to 5.8% — a difference that is much too small to be distinguishable from zero in our sample.²² How shall we make use of our matched sets and our 4 year sample window in such a way that we discover whether any of these 9 safety based withdrawals occurred because of PDUFA I?

We have a sample and an organization of the data which are as similar as possible to a randomized experiment. Since we spent so much effort to choose and organize our data to approximate

²²A fisher’s exact test for $A = 0$ yielded a two-sided $p = 1$.

characteristics of an experiment, it seems natural to estimate treatment effects as if we had an experiment. What does it mean to estimate an effect “as if” we actually had run an experiment?

First, it means that we will make the strong ignorability assumption. Second, it means that we will base our probabilistic inference as much as possible on aspects of the design and as little as possible on modeling assumptions.²³ Random assignment (when it occurs) provides, in Fisher’s words, a “reasoned basis” for inference. That is, if we know how Z was assigned (conditional on covariates), then we can make inferences about the effects of Z on y without having to make assumptions about the distribution of y (Rosenbaum, 2002*b*, Chapter 2). We titled this section, “permutation based inference” and not “randomization based inference”, however, because we do not have randomization. However, even without randomization, we can act as if we had it, and then can (we hope) reap the benefits of inference that requires very few assumptions.²⁴

In our particular case, given our definition of a non-constant, but additive, causal effect τ_i and our estimand $A = \sum_i Z_i \tau_i$, we will use the method of attributable effects (Rosenbaum, 2002*a*, 2001).

5.1 Attributable Effects Briefly Described

Given our matched sets, we can easily test the hypothesis that $A = 0$ (that is, the hypothesis of no treatment effect) under the assumptions of ignorability and SUTVA conditional on our matched sets, and also assuming that the number of outcomes observed are fixed. This test well known as the Mantel-Haenszel (MH) test for $2 \times 2 \times S$ tables (Mantel, 1963). In our case, $S = 70$ (the matched sets) and the 2×2 table is treatment (pre- versus post-PDUFA I submission within the 4 year window) by outcome (no withdrawal, withdrawal). The null hypothesis for this test is that $A = 0$ — no unit in the treatment group would have had any different withdrawal behavior in the control group. The p-value for this test is shown in the second column of Table 3 on the following page under the column for $A = 0$. Clearly, the data do not provide much evidence against this particular null ($p = .84$). The results for the matched sets without prognostic scores are show in the bottom half of the table, but the results are substantively the same.

What if we wanted to move beyond merely testing the null hypothesis of no effect, but we wanted a confidence interval for A ? We can easily generate such a confidence interval by repeatedly testing different hypotheses about A . By definition, a confidence interval of level $100(1 - \alpha)$ is the set of hypothesis accepted at level α . This method of “inversion” of a hypothesis test to produce a confidence interval is well known.²⁵

In this case, the mechanics of the test involve adjusting the observed outcomes to reflect the null hypothesis, and then testing this hypothesis. For example, say we want to test whether $A = 1$. If A truly were 1, then one of the 9 safety based withdrawals which occurred after PDUFA I would not have occurred before PDUFA I. If this truly is the case, then if we subtract 1 from 9 withdrawals and we execute the MH test, we ought to see a large p-value — indicating that we ought to accept this null. If we see a small p-value, then this configuration of outcomes is inconsistent with the null. By testing each A from 0 to 9 and adjusting the outcomes each time to reflect the null hypothesis, we can generate a confidence set for A .

²³We also had to make the SUTVA assumption. This is mostly reasonable in this case since, even competition among firms to get their drugs to the FDA quickly would result in sets with the same kinds of drugs being compared.

²⁴LIST SOME OF THE BENEFITS

²⁵See Rosenbaum (2002*b*, Chapter 2) for more details about this procedure.

One Wrinkle with Attributing Effects One problem occurs when attributing effects to stratified samples. Say, $A = 1$. In our 70 sets, 9 sets contain post-PDUFA drugs which were withdrawn from the market for safety reasons. That is we have 9 sets to which we can attribute an effect of 1. Which one should we choose? In this case, we didn't need to make this choice. We just adjusted each set in turn, administering the MH test each time. This means that a hypothesis of the form $A = 1$ is really a statement about a collection of hypothesis tests. If we accept any one of the tests, then we must accept that $A = 1$. In Table 3, the range of p-values returned for the composite hypothesis $A = 1$ is listed in the third column (hi p=.99 and lo p=.88). This means that of the 9 different MH tests, each of which tested $A = 1$, the p-values ranged from .88 to .99. From the perspective of evidence in favor of the null, these are practically indistinguishable — we accept that it is reasonable to suppose that $A = 1$ ought to be in the confidence set.²⁶

5.2 Results

Table 3 shows the rest of the tests and attributions. Focusing on the rows using the matchings generated with prognostic scores we see that all of the attributions are probable — perhaps with 0,1,2 or 3 attributions more probable than the others. However, any conventional definition for level of a confidence set here would include all of the possible attributions — including zero with high probability.

Propensity Scores Using Prognostic Scores										
A=	0	1	2	3	4	5	6	7	8	9
lo p		0.88	0.76	0.72	0.64	0.48	0.35	0.25	0.18	
hi p	0.84	0.99	1.00	1.00	1.00	0.99	0.83	0.61	0.39	0.14
Propensity Scores Only										
A=	0	1	2	3	4	5	6	7	8	9
lo p		0.61	0.63	0.66	0.66	0.67	0.64	0.59	0.53	
hi p	0.59	0.97	1.00	1.00	0.98	0.91	0.89	1.00	0.94	0.48

Table 3: p-values for different attributions of effect using normal approximations to the exact distribution via `mantelhaen.test` applied to adjusted outcomes.

Table 3 includes the results from two sets of matchings mainly as an attempt to show that the power of tests using the set created with prognostic scores was higher than the power of tests using the set created without prognostic scores. Given this data, this demonstration is hard to see here unless it is hinted at by the more rapid decline in p-values across the attributions in the first set of matchings (from .84 to .18 lo p) versus the second set (.59 to .48 lo p).

²⁶There are more wrinkles here, but none that matter for this particular application since we never have more than about 500 ways to attribute effects.

5.3 Summary

We can't exclude $A = 0$ from our confidence set. Indeed, it is highly probable. This result (although very preliminary) runs counter to the claims of opponents of the user fee acts. For example, see the following from a position paper issued by Public Citizen's Health Research Group:

1. An unprecedented number of drugs withdrawn since the passage of PDUFA

Since the passage of PDUFA in 1992 there have been an unprecedented number of drugs approved and then withdrawn for safety reasons. Nine of the drugs (Baycol, Raplon, Lotronex, Propulsid, Rezulin, Raxar, Duract, Posicor, Redux) approved during the eight-year period from 1993 to 2000 have been withdrawn because of safety concerns. By comparison, only five of the drugs approved during the eight-year period from 1985 through 1992 later had to be withdrawn. In the case of several of the drugs approved since 1992, their approval and the subsequent loss of life they are suspected to have caused was entirely needless. At least five of these nine drugs (Raplon, Raxar, Duract, Posicor, Redux) were approved despite known safety problems and the availability of multiple treatment options in other, older (and safer) drugs approved for the same medical uses. (Lurie and Sasich, N.d.)

Of course, we have more work to do before we have confidence in our results.

5.4 Sensitivity Analysis

TO DO

6 USING REGRESSION AND REPEATED SAMPLING MODEL-BASED HYPOTHESIS TESTING

In an applied statistics paper, it is enough to demonstrate a new method and to argue for its usefulness. Political scientists, however, often need more convincing before they will spend their valuable time learning new (and possibly half-baked) tools. In this section we execute what most political scientists would call a "normal" analysis of this question, and then we will carefully note the assumptions required to believe this estimate and to compare them to the assumptions required for our "not-normal" (but we think, better) mode of analysis.

[TO DO]

6.1 Assumptions Required To Believe This Estimate

[TO DO]

6.2 Comparing the Linear Model and Model-Based Inference Workflow to the Matching and Permutation Based Workflow

[TO DO]

7 DISCUSSION

7.1 Problems

7.1.1 Do we have the wrong counterfactual?

In contrast to a linear modelling approach, the method of analysis that we proposed here is very explicit about what is being compared to what — both at the stage of observation choice and adjustment and also at the stage of estimation of effects. The “treatment” group is clear: those drugs submitted after the implementation of PDUFA I. The definition of the “control” group or the observations representing the counterfactual condition is less clear. From the stand point of advocacy groups like Public Citizen, the drugs that were withdrawn for safety reasons *would not have been approved* had they been submitted pre-PDUFA. However, we don’t observe any drugs that were submitted after Stage III trials and which were subsequently not approved. In fact, this outcome occurs very rarely if at all — it is much more common for a company to have disappointing Stage III (or Stage II) results and to merely not submit the drug at all (but perhaps to try more trials after “fixing” the drug somehow).

The counterfactual that we test here states that these drugs would still have been approved, but they would not have had the safety problems that would have lead to withdrawal from the market place. Is this credible? It does not exactly address the public policy counterfactual (which is unobserved and pretty much unobservable). But is it close enough?

[add discussion]

Note that by not releasing information about failed approval attempts, the FDA effectively prevents evaluation of itself. If we don’t know which “bad” drugs are being prevented from going to market, we cannot know whether the FDA is doing its job as an arbiter of safety and efficacy. So, this lack of data poses a public policy conundrum in and of itself — how can we evaluate the FDA if the most direct way of doing so would harm the drug companies who are paying for the operation of the FDA?

What other approaches might we try?

Use the Clinical Trials Pipeline Look at the clinical trials pipeline. The idea here would be that those drugs which companies think would be marginal (worth marketing but possible safety concerns) would not have been submitted under the purportedly more rigorous pre-PDUFA FDA but would have been submitted after PDUFA.

[elaborate]

Focus on Black Box Warnings and Adverse Events Another way around the fact that we don’t observe non-approved drugs is to notice that there may be other ways to be approved, but not withdrawn. Namely, one thing that the FDA does during reviews is to decide on black box warnings among other safety enhancing protocols to be applied to the marketing and use of the drugs they approve. So, another implication of the “Public Citizen theory” (stating that the current FDA is causing drug safety problems) is that these drugs, if submitted a few years earlier, would still have been approved, but would just have had more black box warnings — or if not those particular warnings, some other communication or regulation on the part of the FDA cautioning against particularly dangerous (unnecessarily dangerous) uses of the drug. And, the withdrawals that we observe that could be due to the new regulatory regime are in fact caused

not by the approval of blatantly dangerous drugs, but by a lack of caution about the specifics of marketing and using drugs which are dangerous but necessary (less dangerous and more life-saving/life-enhancing than alternative drugs).

That is, "withdrawal" captures a lot of different activity on the part of the FDA — analysis of Adverse Event Reports, assessment of alternative therapies, other regulation of the prescription and use and warnings associated with drugs, etc... Some of that activity could occur more (or less) after PDUFA even among approved drugs to cause differences in safety based withdrawals.

So, this means that we can use the data that we have, which consists only of approved drugs. And, if we found that there were more black box warnings on drugs that are similar to the post-PDUFA withdrawn drugs, but drugs that were not withdrawn than on the drugs which were withdrawn post-PDUFA, then we'd have some evidence that this cautionary post-marketing activity on the part of the FDA was part of the process by which the FDA can enhance or diminish safety.

[other problems and caveats]

7.2 Conclusion

We began this paper with two questions: (1), "Have the imposition of deadlines on FDA review by the Prescription User Fee Acts (PDUFA)s caused drug safety problems?" (2) "In an observational study with discontinuities, how can we use as much information as possible (in order to enhance precision) while avoiding bias and fragility of results as much as possible?"

Our answer to the first question is, for now, "no". At least in so far as the user fee act is PDUFA I and drug safety problems are withdrawals from the global market. We will be adding analysis of the PDUFA II and III treatment groups as well as Adverse Event Reports as another indicator of drug safety to our analyses soon.

Our answer to the second question was: (1) choose a set of drugs to compare based on similarity on observed covariates within as small a temporal window around the discontinuity induced by the law change as possible; (2) organize this set of drugs into sets within which the drugs would be as homogeneous as possible in terms of covariates and also in terms of potential to response to the control condition; (3) estimate treatment effects using only assumptions about treatment assignment (plus ignorability and SUTVA as all such studies are required to do) using attributable effects.

Have these set of techniques been fruitful? Especially more fruitful than other more common techniques (say, by merely using a logistic regression or even a logistic regression with fixed effects for matched set membership)? It is possible that such techniques might provide more precise estimates of (no effect) that we have shown here. We are sure, however, that our techniques require fewer and less onerous assumptions from the analyst than other approaches.

REFERENCES

- Achen, Christopher H. 1982. *Interpreting and Using Regression*. Newbury Park, CA: Sage.
- Bowers, Jake and Ben B. Hansen. 2005. "Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign: An Application of Randomization Inference Using Full Matching." Presented at annual meeting of the Political Methodology Section of the American Political Science Association.
URL: <http://www-personal.umich.edu/jwbowers/PAPERS/bhPolmeth2005July.pdf>
- Bowers, Jake and Ben B. Hansen. 2006a. Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign. Technical Report 448 Statistics Department, University of Michigan.
- Bowers, Jake and Ben Hansen. 2006b. "Are In-Person Get-Out-The-Vote Efforts A Civic Good? Or a Civic Bad? or Who Can be Mobilized? Attributing Differential Treatment Effects by Group in a Field Experiment of Turnout." Unpublished manuscript.
- Bowers, Jake and Ben Hansen. 2006c. *RItools: Randomization Inference Tools*. 0.12 ed.
URL: <http://www.umich.edu/jwbowers/RItools.html>
- Campbell, D.T. and J.C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin.
- Carpenter, Daniel, Jake Bowers, Justin Grimmer, Susan Moffitt, Clayton Nall and Evan James Zucker. 2007. "Deadline Effects in Drug Regulatory Review." Robert Wood Johnson Health Policy Scholars Working Paper # 35.
- Carpenter, Daniel and Justin Grimmer. 2007. "Dynamic Stochastic Learning in Regulatory Optimization: Deadlines and Error with Continuous and Discrete Evidence." manuscript, Department of Government, Harvard University.
- Carpenter, Daniel and Michael M. Ting. 2007. "Regulatory Errors under Endogenous Agendas." Manuscript, Harvard University. Earlier distributed as "Regulatory Errors under Two-Sided Uncertainty," RWJF Scholars in Health Policy Working Paper.
- Firth, David. 2006. *brlr: Bias-reduced logistic regression*. R package version 0.8-8.
URL: <http://www.warwick.ac.uk/go/dfirth>
- Grabowski, Henry and Y. Richard Wang. 2006. "Determinants of New Drug Safety in the United States, 1993-2003." Unpublished manuscript.
- Hansen, Ben B. 2004. "Full matching in an observational study of coaching for the SAT." *Journal of the American Statistical Association* 99(467):609-618.
- Hansen, Ben B. 2006a. Appraising Covariate Balance after Assignment to Treatment by Groups. Technical Report 436 University of Michigan Statistics Department: .
- Hansen, Ben B. 2006b. Bias Reduction in Observational Studies via Prognosis Scores. Technical Report 441 Statistics Department, University of Michigan.
- Hansen, Ben B. 2007. "The Prognostic Analogue of the Propensity Score." Unpublished manuscript.

- Hansen, Ben B. and Stephanie Olsen Klopfer. 2005. Optimal full matching and related designs via network flows. Technical Report 416 Statistics Department, University of Michigan.
- Imai, Kosuke, Gary King and Elizabeth A. Stuart. 2006. "The Balance Test Fallacy in Matching Methods for Causal Inference." Unpublished manuscript.
- Lurie, Peter and Larry Sasich. N.d. "Problems in the New Drug Approval Process: A Case Study of Lotronex."
URL: <http://www.citizen.org/congress/reform/drugsafety/articles.cfm?ID=7442>
- Mantel, Nathan. 1963. "Chi-Square Tests with One Degree of Freedom; Extensions of the Mantel- Haenszel Procedure." *Journal of the American Statistical Association* 58:690–700.
- Neyman, J. 1990. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." *Statistical Science* 5:463–480. reprint. Transl. by Dabrowska and Speed.
- Olson, Mary K. 2002. "Pharmaceutical Policy Change and the Safety of New Drugs." *Journal of Law and Economics* XLV:615–642.
- Olson, Mary K. 2004. "Are novel drugs more risky for patients than less novel drugs?" *Journal of Health Economics* 23:1135–1158.
- Rosenbaum, Paul R. 2001. "Effects Attributable to Treatment: Inference in Experiments and Observational Studies with a Discrete Pivot." *Biometrika* 88(1):219–231.
- Rosenbaum, Paul R. 2002a. "Attributing Effects to Treatment in Matched Observational Studies." *Journal of the American Statistical Association* 97:183–192.
- Rosenbaum, Paul R. 2002b. *Observational Studies*. Springer.
- Rosenbaum, Paul R. and Donald B. Rubin. 1983a. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.
- Rosenbaum, Paul R. and Donald B. Rubin. 1983b. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1):41–55.
- Rosenbaum, Paul R. and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79(387):516–524.
- Rubin, D.B. 1974. "Estimating the Causal Effects of Treatments in Randomized and Nonrandomized Studies." *J. Educ. Psych.* 66:688–701.
- Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100:322–331.
- Sekhon, Jasjeet S. 2006. "Alternative Balance Metrics for Bias Reduction in Matching Methods for Causal Inference." Unpublished manuscript.
- Shadish, William R., Thomas D. Cook and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.