Biostatistics (2001), **2**, 2, *pp.* 217–232
Printed in Great Britain

Matching and thick description in an observational study of mortality after surgery

PAUL R. ROSENBAUM*, JEFFREY H. SILBER
Departments of Statistics and Pediatrics, University of Pennsylvania, Philadelphia, PA 19104-6302, USA
Email: rosenbaum@stat.wharton.upenn.edu

SUMMARY

Multivariate matching permits the construction of matched pairs or matched sets that balance large numbers of observed covariates. Unlike model-based adjustments, in matching a patient remains intact as a single patient, and may be scrutinized as an individual and thickly described. A thick description entails a detailed, perhaps narrative, account of a patient's care, for instance, the account one might find in the 'Case Reports from the Massachusetts General Hospital' as published in the *New England Journal of Medicine*. While discussing certain principles of thick description, we illustrate using data from the pilot for a case-control study of the causes of death following surgery. Matching is based on billing data from Medicare, and the medical charts of matched pairs are then abstracted. In the pilot, we matched cases and controls in one hospital, located and scrutinized their medical charts. As a consequence, we corrected our misinterpretations of aspects of Medicare billing data, thereby improving the matching for the full study. Also, looking at charts suggested topics for investigation and helped us understand the types of information we might reliably find in charts, and this reshaped our plans for chart abstraction. Our central claim is that, unlike other methods of adjustment, matching facilitates thick description of a handful of cases, and such scrutiny of cases benefits statistical studies at several stages. Thick description of a few matched cases is used repeatedly to improve the matching and to design further data collection for the matched sample. Thick description aids matching by providing close examination of what the matching has actually accomplished, an examination that uses much more information than is available for use in matching. Matching aids thick description by placing side by side two patients who are fairly comparable, so that a thick description of them may usefully be performed.

Keywords: Ethnography; Full matching; Multivariate matching; Qualitative methods; Thick description.

1. DYING AFTER SURGERY; REVIEW; OUTLINE

1.1 *Introduction: thick description in statistics*

Do measurements recorded in a computer file mean what they appear to mean? Or would direct contact with the people described in the file reveal that the records mean something else entirely? Would people who appear comparable in computer records still appear comparable if observed directly? Or do they differ in important ways that have not been recorded? Would mechanisms theorized to cause certain reactions actually be visible if those reactions were closely scrutinized? Or would close inspection of a few cases show that the mechanisms did not operate in these cases?

*To whom correspondence should be addressed.

Questions of this sort can be addressed with basic applications of ethnographic techniques that compare quantitative measurements of events to a more detailed, more individual, more narrative, thicker description of those same events. Our purpose in this paper is to link the simplest forms of ethnographic or thick description and a particular statistical method which aids rather than inhibits such description, namely multivariate matching. We do this with reference to our ongoing study of mortality following surgery. This introductory section is organized as follows. Our study of surgical mortality is briefly described in Section 1.2, followed by short reviews of matching in Section 1.3 and of thick description in Section 1.4. We then discuss the link between matching and thick description in Section 1.5, and outline the remainder of the paper in Section 1.6.

1.2 *A case-control study of mortality after surgery*

We recently concluded the pilot stage of a case-control study of the causes of death following surgery among Medicare patients in Pennsylvania, and we will soon begin chart abstraction for the main study. Building upon earlier work (Silber *et al.*, 1992, 1995a,b, 1997a; Silber and Rosenbaum, 1997b; Silber *et al.*, 1999) with medical records, the pilot stage closely examined data from one hospital, whereas the main study will examine a random sample of Medicare deaths following surgery throughout Pennsylvania. In the main study, we will match 830 patients who died following surgery to patients ostensibly similar in health upon admission to the hospital, but who did not die. The matching is based on computerized Medicare billing data, and for the matched pairs, we will abstract the hospital records in the hope of finding preventable causes of death. Most patients treated surgically are comparatively healthy and do not die, and even among the severely ill, most surgical patients do not die during the few months after surgery. Given this, a case-control study is needed, because it would be prohibitively expensive and extremely wasteful to abstract charts for the vast majority of healthy patients who do not die. Matched sampling is needed to compare deaths to patients who were, upon admission, severely ill and at high risk of death. After hospital charts are abstracted, models will be used to control for biases that are visible in the charts but which were not visible in the Medicare billing data.

Medicare billing data records medical procedures performed, principal and secondary diagnoses as International Classification of Diseases (ICD-9) codes, basic demographics, and limited outcome information including mortality. Data on a single patient over time can be linked, so similar information may be obtained for a sequence of hospital and outpatient visits. Clinical detail, such a record of vital signs, drugs received, and lab results, is not available from Medicare and must, instead, be obtained by chart abstraction. Since payments to hospitals are tied to medical bills, there are significant incentives for fraud in Medicare's bill data; however, such fraud is a felony, and significant penalties are regularly imposed.

This case-control study is unusual. Typically, case-control studies focus on disease etiology in natural populations, for instance, the effects of occupational or environmental hazards, such as asbestos or radon, or the effects of personal habits, such as smoking or diet. The question is: what unintended exposures or treatments caused the disease? Typically, studies of medical and surgical treatments focus on particular treatments that are deliberately applied; for instance, the effects of coronary bypass surgery versus drug treatments for coronary disease. The question is: which of several intended treatments will produce the best outcomes for specific types of patients? Both types of investigation are important, but there is a substantial, largely unexplored, region between them. There are many alterable aspects of health care, particularly the care surrounding surgery, which are not intended treatments, but which may have substantial effects on patient outcomes. These aspects are part of health care, and in this sense resemble bypass surgery, but they are not applied or withheld from patients in a deliberate way, with a goal in mind, and in this sense they resemble occupational or environmental hazards. These unintended aspects of health care are quite varied. Some involve the qualifications of the staff who provide the care. Was the anaesthetic

given by an anaesthesiologist or by a nurse? Was the surgeon board certified? What fraction of nurses are registered nurses? Are there residents in the hospital? In Silber *et al.* (1995a,b), some of the stronger associations between hospital characteristics and patient outcomes were found for staff qualifications. Other aspects of health care involve responses to complications. Typically, before a patient dies, the patient experiences one or more complications. A death following a complication is called a failure-to-rescue (Silber *et al.*, 1992). Did the hospital respond to the complication in a timely and appropriate way? When a drug was ordered, was the order filled promptly? Are patients properly monitored? Were intensive care units used when needed? Were mistakes made? Still another aspect of health care involves the unintended or unanticipated side effects of intended treatments, such as side effects of specific drugs or drug interactions or of specific medical tests. The mortality rate in the 30 or 60 days following surgery is high compared to the mortality rate in the entire Medicare population, and some part of that mortality rate may be preventable by altering health care. Because our interest is in the care provided by the hospital, not the cause of the disease, we would like to compare patients who were comparable when admitted to the hospital.

As part of the pilot stage, we matched 38 deaths following surgery in one hospital to 38 similar patients who did not die. These were all available deaths in this hospital in 1994. A variant form of the matching in the pilot stage with these data is described by Ming and Rosenbaum (2000). This pilot data will not be included in the final study which will use data from more recent years. Therefore, insights gleaned from the pilot data may be confirmed with new data that did not contribute to these insights. After matching, we pulled and read the medical charts for selected matched patients. This led to a substantially more accurate understanding of the relationship between the Medicare billing data available for matching and the hospital charts, and a substantially more realistic image of what we would find in charts when they are abstracted. Using this revised understanding, we rematched the 38 deaths several times, reading charts after each match, continuing until we were satisfied that we were getting everything we could from the Medicare billing data. In this paper, we discuss this informal, but nonetheless important, process of comparing computer records with a thicker text describing the same events, and the ways such a comparison improves a statistical analysis. Moreover, we link the particulars of this example to issues and principles that are applicable in general.

1.3 *Review: matching in observational studies*

An *observational study* compares treated and control groups that were not formed by random assignment in an effort to draw inferences about the effects caused by the treatment (Cochran, 1965; Rubin, 1974; Rosenbaum, 1995, 1999). Because treatments were not randomly assigned, treated and control subjects are often not comparable prior to treatment, so differing outcomes may reflect these pretreatment differences rather than effects of the treatment. Pretreatment differences in observed and accurately measured covariates constitute an *overt bias*—such a bias is visible in the data at hand and is removed by adjustments. Relevant pretreatment differences that are not observed constitute a *hidden bias*—they cannot be removed by adjustments, and must be addressed by other methods, such as sensitivity analysis (Rosenbaum, 1991a, 1995, 1999, Section 4). The distinction between hidden and overt bias is relative to the data at hand: a bias may be hidden in the Medicare billing data and overt in the more detailed medical chart.

Adjustments for overt biases may be done using several techniques, alone or in combination. In *matching*, a treated subject is compared to one or more control subjects who appeared similar in terms of measured covariates (Rubin, 1973; Smith *et al.*, 1977; Rosenbaum and Rubin, 1985; Mosteller, 1996; Ming and Rosenbaum, 2000). In *stratification* or subclassification, groups of treated subjects are compared to ostensibly similar groups of control subjects and the results are combined across these apparently balanced strata (Cochran, 1968; Rosenbaum and Rubin, 1984). The optimal form of a stratification—the

form that makes subjects in the same stratum as similar as possible—is a variant form of matching, called *full matching*, in which a treated subject may be matched to several controls or a control may be matched to several treated subjects (Rosenbaum, 1991b). Full matching often removes much more bias than other matching methods (Gu and Rosenbaum, 1993). Moreover, when costly additional data collection is not required, full matching permits all available controls to be used. In *model-based adjustments*, such as covariance adjustment, outcomes are modelled in terms of observed covariates and assigned treatments, leading to adjusted estimates of treatment effects as coefficients of the model. The methods may be applied in combination. For instance, Rubin's (1979) simulation shows that covariance adjustment of matched pair differences can be more efficient than matching alone and more robust to model misspecification than covariance adjustment alone. Covariance adjustment of matched pairs is illustrated in Rosenbaum (1986) in a study of US high school dropouts. The hazards of using model-based adjustments unaided by matching or stratification are discussed by Dehejia and Wahba (1999).

Progress in matching and stratifying on many covariates simultaneously has been possible by shifting attention toward *balanced or comparable matched groups*, and away from perfectly homogeneous matched sets or strata. Although homogeneous matched sets or strata are desirable, it is often difficult to find subjects with identical values of many covariates, whereas it is much less difficult to assemble treated and control groups that look comparable in aggregate. The *propensity score* is a device that permits matching or stratification on many covariates at once, yielding matched sets or strata that balance, but are not necessarily similar in observed covariates (Rosenbaum and Rubin, 1983; Joffe and Rosenbaum, 1999). The propensity score is the conditional probability of treatment given observed covariates, and matching or stratifying on an estimate of the propensity score can balance many covariates (Rosenbaum and Rubin, 1984, 1985). Simulation suggests that matching on propensity scores is much better than competing methods when there are many covariates (Gu and Rosenbaum, 1993) and Dehejia and Wahba (1999) reach a similar conclusion by contrasting several analyses of one empirical study. Because closely matched sets are desirable, even though they are less attainable than balanced groups, modern matching methods insist on covariate balance for groups as a whole, and then attempt to maximize *proximity* within matched sets by minimizing a distance that emphasizes the propensity score (Rosenbaum and Rubin, 1985; Gu and Rosenbaum, 1993). Fast algorithms exist for finding the optimal match (Rosenbaum, 1989; Gu and Rosenbaum, 1993; Ming and Rosenbaum, 2000) and these are available through the computer package SAS (Bergstralh *et al.*, 1996).

The biases matching and adjustments are intended to remove do not diminish with increasing sample size, that is, they are O(1) biases, whereas estimators typically have variances that diminish with increasing sample size N: that is, their variances are $O(\frac{1}{N})$. For this reason, if any bias remains, it quickly dominates the mean square error as the sample size increases, so removing biases of O(1) becomes infinitely more important than statistical efficiency. For instance, Cochran (1965, §2.1) describes 'protecting against bias' as the 'primary role' of matching, subclassification and adjustments in observational studies. Cochran (1965, §3.1) and Breslow and Day (1980, p. 107) caution against ignoring a pretreatment covariate based on the outcome of a test of association. In some instances, there are either no observed pretreatment covariates at all or else there are just a few coarse covariates, so all of the bias due to observed covariates can be removed, and in this case, statistical efficiency becomes the focus, as it is in randomized experiments. Chase (1966) considered the situation in which there is no covariate and no bias, comparing matching at random to an unmatched analysis, reporting, in this special case, very similar Pitman, Bahadur and Hodges–Lehmann efficiencies for matched and unmatched analyses. Brookmeyer *et al.* (1986) considered the case of a few coarse covariates and compared a matched analysis to a stratified analysis, finding comparable efficiencies for small odds ratios less than two, and higher efficiency for the stratified analysis for large odds ratios of five. Two decades ago, the cost of constructing a matched sample by hand was significant (Thompson *et al.*, 1982), but today, with the easy availability of SAS macros for multivariate matching (Bergstralh *et al.*, 1996), constructing a multivariate matched sample

from computer records costs little more than fitting a logit model. In our study, there are many covariates, matching removes only part of the overt bias, and so bias reduction, not efficiency, remains the focus.

1.4 *Review: thick description*

1.4.1 *Origin of the term 'thick description'.* The term 'thick description' was introduced by Ryle (1971), a philosopher, and was made popular by Geertz (1973), an anthropologist, in a discussion of ethnographic methods. In extremely general terms, a thick description of human events and behaviour is one that retains and is faithful to the meanings which that behaviour has for the people involved. That explanation is, however, too vague to indicate what is at issue. For example, an advocate of thick description would be critical of a behaviourist description of behaviour: that is, a description of behaviour exclusively in terms of overt physical events, with the intentions, purposes, beliefs, conjectures and thoughts of the participants removed as inappropriate for scientific study. An advocate of thick description would probably avoid describing thoughts and behaviour in terms derived from a general models of behaviour, whether that model speaks of optimizing utility or unconscious wishes. Instead, the description would faithfully use terms and concepts that the individuals being described would recognize and judge appropriate. An advocate of thick description would also be critical of any insistence that an interaction between people must have a single meaning or interpretation, rather than multiple, perhaps conflicting, perhaps unstable, interpretations held by the several different participants, who may have different or changing purposes or intentions. An advocate of thick description would claim that brief, numerical descriptions of behaviour often distort meaning, and would therefore favour a rich, narrative account, perhaps accompanied by photographs, film, taped conversations, an account that makes the reader feel acquainted with the individuals described. Most simply, an advocate of thick description would be sceptical of a claim that a quantitative data file is always a complete and accurate record of the course of events. Anyone who has ever tried to use medical billing data or hospital charts to understand human health would have sympathy with several of these concerns. The specific concern that available data may omit important measurements is closely connected to the concern about hidden biases which are central in observational studies.

Ryle, the philosopher, attempted to clarify the concept of 'thick description' in the way philosophers clarify concepts, namely using crisp illustrations drawn from common experience, placing little emphasis on their relevance in scientific explanations. To given the flavour of Ryle's (1971, p. 479) examples,

> [consider] ... the notion of waiting—waiting for a train perhaps ... The 'thick' description of what I am doing on the platform requires mention of my should-be train-catching. Here there is [nothing] in particular that I must be positively doing in order to qualify as waiting. I may sit or stand or stroll, smoke or tackle a crossword puzzle, chat or hum or keep quiet. All that is required is that I do not do anything or go anywhere or remain anywhere that will prevent me catching the train. Waiting is abstaining from doings that conflict with the objective.

'Waiting' is readily characterized in terms of purposes, and poorly in terms of specific behaviours. Indeed, although one might deduce from overt behaviour that someone is waiting for a train, that behaviour does not become 'waiting for a train' until its purpose is recognized. Purpose organizes behaviour, and our attribution of purpose to someone organizes, perhaps correctly, our perception of that person's behaviour. It is not possible to give a coherent reading of a hospital chart except by attributing, perhaps correctly, purposes to the medical staff whose actions are recorded.

The relevance of thick description to empirical research is emphasized by the sociologist Becker (1996, p. 58):

> ... if we don't find out from people what meanings they are actually giving to things, we will still talk about those meanings. In that case, we will, of necessity, invent them, reasoning that the people we are writing about must have meant this or that, or they would not have done the things they did. But it is inevitably epistemologically dangerous to guess at what could be observed directly. The danger is that we will guess wrong, that what looks reasonable to us will not be what looked reasonable to them. This happens all the time, largely because we are not those people and do not live in their circumstances.

What does a thick description hope to describe? Shweder (1996, p. 177) writes:

> ... I actually think the well-publicized tension between quantitative and qualitative approach has a greater ring of truth when formulated as a problem in ontology rather than as a problem in method (or epistemology) ... quantitative research (with its methodological emphasis on pointing, sampling, counting, measuring, calculating, and abstracting) is premised on the notion that the subjective involves illusions that should be rejected. The basic idea is that it is only when all subjectivity has been subtracted from the world that the really real world remains. And what remains that is really real is the world of *quanta* ... In contrast, qualitative research (with its procedural emphasis on empathy, interpretation, thematization/enplotment, narration, contextualization, and exemplification/concreteness/substance) is premised on the notion that the objective conception of the real world is partial or incomplete. The basic idea is that one of the very important things left out of the real world by the objective conception are *qualia*. Think of *qualia* as things that can only be understood by reference to what they mean, signify, or imply ....

Geertz wrote:

> ... ethnography is thick description ... Doing ethnography is like trying to read (in the sense of 'construct a reading of') a manuscript—foreign, faded, full of ellipses, incoherencies, suspicious emendations, and tendentious commentaries, but written not in [... words ...] but in transient examples of shaped behaviour. (1973, p. 9) ... [The aim of ethnography is] to render obscure matters intelligibly by providing them with an informing context. (1983, p. 152)

1.4.2 *Role of thick description in large quantitative studies.* Geertz's phrase—'render obscure matters intelligibly by providing them with an informing context'—may be the best definition of thick description, at least as it relates to the statistical analysis of large quantities of data. In this one specific context, thick description renders obscure aspects of the quantitative data more intelligible by providing an informing context for a handful of cases. That context can be derived from various sources. Most simply, the context may come from reading the entire hospital chart as a narrative document, as distinct from a brief, coded abstraction of a hospital 'front sheet' or billing summary, used for administrative purposes, that may contain less than 100 variables. The chart of a patient who dies in the hospital is often a folder that is several inches thick. Such a reading may be aided by the insights of surgeons and critical care specialists. The context could also come from discussions with surgeons or patients involved in specific cases, or from direct observation of similar cases and their subsequent transcription into hospital chart, Medicare bills, and chart abstractions. In this one specific context, the goal is to make better and more accurate use of the recorded data by better understanding of the relationship between the data and the actual actions, events, and interactions that the data describe.

In addition to informing quantitative analysis, a thick description can yield a second, independent analysis of the same topic but from a different perspective. The matching we describe may help to

coordinate these two analyses. We do not illustrate thick description in this larger sense, nor are we qualified to do so. For two excellent, extensive examples of thick description in medical contexts, see Bosk's (1981) study of surgical failure and Estroff's (1985) study of outpatient psychiatric patients. For an application of qualitative methods in psychology, see Katz (1999). A definition of thick description does less to define the term than a good example, and these are three good examples. Helper (2000) discusses related issues in economics. Emerson (1981) surveys the methodology of thick description with extensive references to a large literature.

1.5 *A key concept: the link between matching and thick description*

Matching and thick description are linked by a single, simple fact. Unlike model-based adjustments, where patients vanish and are replaced by the coefficients of a model, in matching, ostensibly comparable patients are compared directly, one by one. Modern matching methods involve statistical modelling and combinatorial algorithms, but the end result is a collection of pairs or sets of people who look comparable, at least on average. In matching, people retain their integrity as people, so they can be examined and their stories can be told individually. Matching is a method of adjustment that facilitates, rather than inhibits, thick description. And, as we will show in an example, thick description can substantially improve a matched sample, altering who is matched to whom.

Our proposal for combining matching with thick description has several parts. First, the close scrutiny of a few matched cases is used repeatedly to improve the matching. Second, when the matching appears satisfactory, further scrutiny of matched cases helps to shape the collection of additional data from matched subjects: in our case, the expensive process of chart abstraction. Finally, quantitative conclusions may be supplemented with detailed qualitative descriptions of a small number of cases and their matched controls.

Thick description aids matching by providing close examination of what the matching has actually accomplished, an examination that often goes beyond the quantitative data available for use in matching. In our study, by looking at charts of matched patients, we were able to make more effective use of Medicare billing data in subsequent matching.

Matching aids thick description by placing side by side two patients who are fairly comparable. In the absence of matching, the records of two surgical patients, picked arbitrarily or at random, would likely be so different that a qualitative comparison of these two patients would conclude, simply, that they are too different to be usefully compared. Matching can strengthen a qualitative case study in two ways: first by supplying one or more controls for qualitative comparison, and, second, by embedding the qualitative case-control study in a much larger quantitative study.

Key conclusions are often comparisons—they cannot be reached by looking at an individual. For example, at several stages we concluded that our first definition of a quantitative variable was formally correct but too vague to be useful. This conclusion was reached by comparing two matched subjects who were the same in terms of this variable, but whose charts revealed them to be medically very different. Had we looked at either chart alone, we would have found that the variable correctly described that one chart. It was necessary to compare two charts to notice that the variable failed to capture important aspects that distinguished the charts. Knowing the two charts were different, we reviewed the Medicare billing data to identify variables that would correctly capture the distinctions our initial definition had missed. Matching and thick description combine repeatedly to produce findings that cannot be reached by either method alone.

1.6 *Outline: thick description in matched observational studies*

Thick description is typically applied to a handful of cases, and Section 2 discusses two principles about what can and what cannot be learned from a few cases. In Section 3 we interleave our discussion of principles with their application in our study of mortality after surgery. Section 4 briefly discusses the choice of matched pairs that will be subjected to close scrutiny, and a brief summary is given in Section 5.

2. CASES AS EVIDENCE: TWO PRINCIPLES

2.1 *Cases may refute an 'observation categorical'*

In discussing Popper's (1968) philosophy of science, Quine (1992, p. 13) writes: 'pure observation lends only negative evidence, by refuting an observation categorical that a proposed theory implies.' This compressed statement says, first, that a scientific theory may imply that under observable circumstances C, result R will always be observed to follow (i.e. Quine's 'observation categorical'), and, second, that a few, quite arbitrary, instances (Quine's 'pure observation') in which C arises but R does not follow suffice to refute the theory. This point is also made by Becker (1996, p. 54): ' . . . every general law supposes that the investigation of particular cases would show that law at work.' A universal claim may be refuted by a few particulars.

Scientific theories of human health or behaviour are rarely this specific. For instance, we are quite confident that smoking causes lung cancer, even though not every smoker develops lung cancer. A few cases of smokers who did not develop lung cancer does nothing to refute the general claim. The general claim is about the risk of lung cancer; it is not a observation categorical.

Nonetheless, a statistical investigation of risks will include within it many observation categoricals. One such observation categorical is: the variables we use mean what we think they mean, and not something very different. A patient interpreted to be free of cancer based on Medicare billing data should not be found in the chart to have documented, metastatic, obstructive pancreatic carcinoma. Another is: Matched subjects may not always be very similar, but they are never wildly incomparable. As described in more detail later, examination of charts for a few of our first matched pairs showed both of these general statements to be false, and led us to revise our use of the Medicare data and our matching in important ways. Also, after revising definitions and the matching, our subsequent chart reviews no longer found such problems, suggesting improvements had been made.

The sharp edged logic of observation categoricals and refuting instances is really only part of the story. A few cases violating a general claim will have much more force if they are accompanied by a clear explanation that appears correct on external grounds, and if there is reason to expect further violations for the same reason. For instance, one may discover that a variable means something quite definite, something quite sensible, but not what was anticipated. A few cases may correct 'aspect blindness', discussed in Section 3.2.

In medicine, case studies are seen as a useful complement to quantitative studies. For instance, the *New England Journal of Medicine* routinely publishes reports of large randomized controlled trials adjacent to 'case reports from the Massachusetts General Hospital'. Becker (1970, p. 75) says that 'case studies', though widely used in the social sciences, originated in medicine.

2.2 *Variety of consequences*

Quantitative or statistical evidence is very different from qualitative evidence or thick description; so different, in fact, that as a matter of taste, quantitative researchers often want nothing to do with qualitative evidence, and qualitative researchers often want no part of quantitative evidence. This preference for

one's own kind of evidence hinders effective investigation. In evidence, variety is a strength, not a weakness.

In his book, *Mathematics and Plausible Reasoning*, Polya (1954) discusses 'patterns of plausible reasoning'. These patterns are familiar from both mathematical heuristics and daily life, and Polya relates them to formal Bayesian reasoning. Quoting Polya (1954, p. 6), one pattern he favours is the following, where three premises are separated by a line from a conclusion, which in turn is followed by Polya's discussion:

(i) A implies B_{n+1}
(ii) B_{n+1} very different from the formerly verified consequences $B_1, B_2, \ldots, B_n$ of A
(iii) B_{n+1} true

A much more credible

> This pattern adds a qualification to the fundamental inductive pattern. Certainly the verification of any consequence strengthens our belief in a conjecture. Yet the verification of certain consequences strengthens our belief more and that of others strengthens it less. The pattern just given brings to our attention a circumstance which has a great influence on the strength of inductive evidence: the variety of the consequences tested. The verification of a new consequence counts more if the new consequence differs more from the formerly verified consequences.

Later, Polya goes on to define 'plausible inference' in formal terms. In Polya's sense, an inference is a 'plausible inference' if every Bayesian, regardless of prior beliefs, would agree with the direction of the inference, although different Bayesians, with different priors, might disagree about force of the inference. In particular, he shows that the argument just given is the basis for a 'plausible inference'. Plausible inference is relevant to the choices made during research design, since a choice that strengthens the evidence in the sense of 'plausible inference' is one that strengthens the evidence for every Bayesian regardless of prior beliefs.

Thick description of a few cases is of value in statistics precisely because this type of evidence has so many strengths and weaknesses that are completely different from the strengths and weaknesses of quantitative evidence.

3. Thick description in statistics

3.1 *Avoiding misinterpretations and misreadings of data*

The simplest, though perhaps most important, contribution of thick description to statistical investigations is as a check that quantitative data are being interpreted correctly. The most important improvement we made based on chart reviews was also the simplest: we redefined 'cancer'.

The Medicare billing data we have for all Medicare patients in Pennsylvania were not collected for medical research and are not always easy to interpret for that purpose. In particular, the billing data do not sharply distinguish medical problems a patient had before coming to the hospital from problems the patient developed during the hospital stay, even those these two types of problems play a very different role in studying the care provided in the hospital. For instance, a patient may come to the hospital with congestive heart failure, or the patient may develop congestive heart failure following surgery, and these are medically very different things, even though Medicare might be billed in the same way. In particular, the 'Universal Bill UB-92' records principal and secondary diagnoses as International Classification of

Diseases (ICD-9) codes, but it does not always distinguish conditions that develop during the hospital stay from conditions the patient brought to the hospital.

There is a common view that strongly discourages matching on any variable whose status is in the least bit questionable. This conventional view is that so-called 'over-matching' is more serious than 'under-matching', because matching on a variable in the design forces one to adjust in the analysis, whereas leaving a variable unmatched permits but does not require adjustments in the analysis. The conventional view is correct in part, but it oversimplifies certain issues (Rosenbaum, 1984; Robins, 1989), and it may overstate the ability of models to correct for very large differences between groups (Dehejia and Wahba, 1999). Following this conventional view somewhat mechanically at first, we initially used as matching variables only variables that rather clearly described what was known about the patient prior to admission to the hospital. We judged a patient to have a history of a condition if the patient had previous bills for this condition, either hospital bills or outpatient bills, prior to the current hospitalization. This turned out to be quite wrong.

In our first matched sample, we matched patients with a history of cancer who died following surgery to ostensibly similar patients with a history of cancer who did not die. In most such pairs of charts that we examined, the patient who died turned out to have an abdominal cancer that might quickly kill the patient, quite apart from the events surrounding surgery, whereas the surviving patient was often fighting a long-standing and relatively quiet battle with breast or prostate cancer. Clearly, these pairs were not comparable.

Everything we know about cancer tells us that a hospital cannot cause measurable amounts of cancer during the course of a one or two week hospital stay. If a patient has measurable cancer during the hospital stay, then the patient brought that cancer to the hospital, no matter when it was first recorded. Therefore, we redefined cancer to include cancer first recorded as a condition in the current hospital stay: for instance, cancer discovered during surgery. Moreover, this permitted us to distinguish different types of cancer in matching, and to match on the patient's current cancer or cancers. In our later matched samples, deaths with serious, life threatening cancers were matched to survivors with similar cancers. We made similar adjustments in the definitions of a number of chronic disorders, like diabetes, that could not have begun during the hospital stay, no matter when they were first noted and recorded during the hospital stay.

Congestive heart failure (CHF) was more difficult. If a patient had a diagnosis of CHF during the current admission, that might reflect a condition brought to the hospital—called a *comorbidity*—or it might reflect a condition developed during the hospital stay—called a *complication*. The Medicare billing data alone simply cannot perfectly distinguish these two cases. Nonetheless, our examination of charts changed the way we used the data about CHF, and our revised definition appeared to work better in the charts we examined. A patient who had CHF during the current admission but no previous record of CHF was defined to have a complication of CHF. A patient who had a history of CHF and a current ICD-9 code for acute pulmonary edema was called both a comorbidity and a complication. However, a patient who had other current ICD-9 codes indicating current CHF together with a history of CHF from bills for previous treatment was defined to have a comorbidity of CHF. Imperfect though this is, it seemed to more closely approximate what we found in the charts.

Notice carefully that the problems with our first matches were not technical. Our analyses showed the matched patients were close on the variables used for matching. However, we were matching on the wrong variables because we misunderstood what the variables meant.

Notice also that the problems with the definition of cancer became apparent not from single charts but from the comparison of two matched charts. Each patient in a pair did indeed have cancer, so the problem with the definition was not obvious from one chart. The problem was obvious from our initial matched pairs of two charts: the death typically had a cancer that presented an immediate danger, whereas the control faced a less immediate risk. An accurate variable, such as the old definition of cancer, may nonetheless be an inadequate variable, in the sense that it fails to capture important distinctions that

become apparent upon closer examination of pairs of charts.

3.2 *Aspect blindness*

The example in Section 3.1 concerned a mistake, that is, an initial understanding of the data that we subsequently rejected as incorrect. Aspect blindness is different. One's past understanding is not rejected, but rather it is seen to be partial and incomplete.

The term 'aspect blindness' is due to Wittgenstein who discussed it in several works: see the selections edited by Anthony Kenny in the chapter 'Aspect and Image' of Wittgenstein (1994). Wittgenstein used a variety of examples of aspect blindness, including drawings—really, optical illusions—which could be 'seen' alternately as different objects with no change in the drawing itself. Wittgenstein wrote:

> The expression of a change of aspect is the expression of a *new* perception and at the same time of the perception's being unchanged. The 'aspect-blind' will have an altogether different relationship to pictures from ours.

Certainly, a central goal in qualitative research is to overcome aspect blindness, to see new aspects of a situation that was, in a sense, already visible. For instance, Blumer (1969) writes:

> The entire act of scientific study is oriented and shaped by the underlying picture of the empirical world that is used. This picture sets the selection and formulation of problems, the determination of what are data, the means to be used in getting data, the kinds of relations sought between data, and the forms in which propositions are cast ... These premises are constituted by the nature given either explicitly or implicitly to the key objects that comprise the picture. The unavoidable task of genuine methodological treatment is to identify and assess these premises ... Because of such a decisive role in scientific inquiry, concepts need especially to be subject to methodological scrutiny. (p. 24) Inside of the 'scientific protocol' one can operate unwittingly with false premises, erroneous problems, distorted data, spurious relations, inaccurate concepts, and unverified interpretations. (p. 29)

In reviewing charts of surgical patients, we were surprised by the number of deaths among patients with serious psychiatric problems. Moreover, our entirely informal impression of their charts suggested a course of treatment that was less organized and goal directed than the treatment of other patients. We became interested in the relationship between psychiatric problems and care following surgery. Do physicians and others react to patient complaints? Do surgical patients with severe psychiatric problems provide a disorganized, perhaps inappropriate, sequence of complaints? Does an alert, focused, polite patient receive better care than a disoriented, enraged, depressed patient? We have no clear opinion or definite answers. However, we added questions to our chart abstraction in the hope of learning about this issue.

3.3 *Definitions*

Our chart abstractions entail numerous definitions. Commonly, one imagines that the world and our perceptions of it come first; that words and their definitions come later. Koch (1999) argues this common view of definition inverts the roles of definitions and perceptions:

> The problem of definition is a special form of the problem of perceptual training. And any definition is at bottom an attempt to guide the addressee toward making a relevant perceptual discrimination. (p. 162) In a loose sense it can be said that words index and, within limits, stabilize discriminal experiences. (p. 160) The problem of definition then becomes the problem of teaching persons how to perceive that constant relational attribute. (p. 164)

Certainly, the definitions involved in data collection in general and chart abstraction in particular are of this form. These definitions are intended to train the abstractors to make certain perceptual distinctions. In numerical data derived from chart abstractions, our definitions determine what we observe. In the quantitative analyses, we have no opportunity to perceive anything we have not previously defined. Our opportunity to check the fitness of the definitions empirically is in the thick description of individual charts.

Definitions shape observations. Goffman (1974, 1986, p. 39) writes:

> . . . observers actively project their frames of reference into the world immediately around them, and one fails to see their so doing only because events ordinarily confirm these projections, causing the assumptions to disappear into the smooth flow of activity.

Bittner and Garfinkel (1967, p. 195) make the same point in the context of interpreting clinical records:

> If the researcher insists that the reporter furnish the information in the way the form provides, he runs the risk of imposing upon the actual events for study a structure that is derived from the features of the reporting rather than from the events themselves.

If definitions shape observations, empirical work must empirically examine definitions. One way to do this is to compare quantitative observations based on these definitions to a thicker, more narrative, less constrained description of the same events. Blumer (1969, p. 24) writes: '. . . methods of study are subservient to [the empirical] world and should be subject to test by it.'

Consider, for example, the definition of the 'timeliness of care'. Our initial conception was to record, as a measure of timeliness of care, the time from diagnosis of each condition to the start of treatment. Examination of charts reveals that this definition works for many conditions but not for some others. In some cases, diagnosis is a gradual process, and the start of treatment may begin when the condition is suspected, but before the diagnosis is confirmed. Timely treatment may mean the time from definitive diagnosis to the start of treatment is a negative time, and it may also mean that some patients are treated for a condition they do not have. If the treatment is mild and the condition is severe, to wait for a definitive diagnosis may be untimely and irresponsible. For other conditions, treatment is an aspect of diagnosis: if antibiotics cure an infection, then the infection is not a viral infection. Definitions must be tested against observations to see if they work.

More fundamental is the definition of a case. We had initially defined a case of mortality following surgery as a death within 30 days of the operation. We do not use in-hospital mortality because elderly patients are sometimes discharged to hospices or other locations not because they are cured but because they are unlikely to benefit from the hospital's services. When we examined matched charts, we discovered one 'control' who never left the hospital and who died a little more than a month after the operation. The chart indicated the patient died of the diseases that brought the patient to the hospital, and that the operation did not save the patient. This patient is better viewed as a case of mortality shortly after surgery than as a control. Among the 38 controls, several others left the hospital, providing no additional chart data, but then died shortly after 30 days from unknown causes. We redefined a case as a death within 60 days of the operation. With this new definition, matched controls typically survived for a long period after their operations.

3.4 *Expert advice*

Surgeons and critical care specialists can provide insights into the care received by a single patient or a pair of ostensibly comparable patients, but they might have little to contribute about estimates obtained from a model. We presented aspects of patient charts to our advisory panel of specialists.

'When a patient arrives at the intensive care unit from surgery', we asked the head of an ICU, 'what measurement of the patient's health and condition do you look at first?' His answer surprised us. 'The time in the operating room,' he said, 'whether the operation took longer than operations of this type usually take.' Time in the operation room—an objective measure, common to all operations, easily obtained, and yet not generally available in administrative data sets, and not generally utilized in models concerning medical outcomes and quality of care. The insights of experts examining a few carefully selected cases can lead to better quantitative data and, as a consequence, to better statistical models.

4. Selecting pairs for close examination

Which pairs should be selected for close scrutiny and thick description? To answer this, one needs to consider how matching algorithms work.

Modern matching methods work first at balancing many observed covariates in treated and control groups, so that the groups appear similar in aggregate prior to treatment. Only as a secondary objective do modern matching methods attempt to form individual pairs that are homogeneous in covariates. The reason is that it is often possible to balance many covariates simultaneously, but it is often difficult to find pairs of individuals who are nearly identical with respect to many covariates: balancing is often practical when close individual pairs are not. For detailed discussion and illustration, see Rosenbaum and Rubin (1985). For instance, one matching method considered there uses calipers on the propensity score to balance covariates, and within calipers, picks individual pairs that are close in terms of the Mahalanobis distance. All pairs are close on the propensity score, thereby balancing covariates, and some are also close on all the individual covariates, as reflected in a small Mahalanobis distance. Gu and Rosenbaum (1993) found this method performed well in simulations.

As a result, the output of a matching algorithm is likely to include balanced matched groups made up of many pairs that are not closely matched, together with a smaller number of pairs that are closely matched. Covariate balance is a property of large groups of people, not of individuals, so it is not a property visible in individual pairs. If, upon examination, the closely matched pairs—say, the pairs with small Mahalanobis distances—appear to be fairly representative of all pairs in terms of covariates, then these pairs are a reasonable choice for closer scrutiny and thicker description. In matching with multiple controls or in full matching, a single matched set will offer several pairs for thick description, and the pair with the smallest distance will be preferred. The hope is that these closely matched pairs will appear comparable when thickly described, and if not, the intention is to redesign the measurement and matching procedures to produce this result on a subsequent attempt. In the final report, narrative accounts of these pairs might accompany the quantitative analysis.

Another issue concerns pairs that are concordant for interesting risk factors. If the risk factor is present for both case and control, or if the risk factor is absent for both case and control, the pair is concordant. If the risk factor is different for case and control, the pair is discordant. In the common methods of analysis, such as McNemar's test and conditional logit regression, concordant pairs do not contribute. Langholz and Borgan (1995) have suggested various strategies for oversampling the informative discordant pairs, a technique they call counter-matching. To the extent that it is possible to identify discordant pairs from the quantitative data, it will typically be more interesting to thickly describe discordant pairs with small Mahalanobis distances.

5. Summary: matching and thick description

We have argued that it is important to compare numerical data to a thicker description of the same events, and that among methods of adjustment for overt biases, matching facilitates rather than inhibits this

process. Such a thick description may improve the interpretation and definition of measures, improve the matching and subsequent data collection for matched subjects, and enrich the presentation of conclusions.

ACKNOWLEDGEMENTS

This work was supported by grant R01-HS9460, 'Surgical Outcome Rates: Identifying Etiologic Factors' from the US Agency for Healthcare Research and Quality, and by a grant from the US National Science Foundation.

REFERENCES

BECKER, H. S. (1958). Problems of inference and proof in participant observation. *American Sociological Review* **23**, 652–660.

BECKER, H. S. (1970). *Sociological Work*. Chicago: Aldine.

BECKER, H. S. (1996). The epistemology of qualitative research. In Jessor, R., Colby, A. and Shweder, R. (eds), *Ethnography and Human Development*, Chicago: University of Chicago Press, pp. 53–72.

BERGSTRALH, E. J., KOSANKE, J. L. AND JACOBSEN, S. L. (1996). Software for optimal matching in observational studies. *Epidemiology* **7**, 331–332. See also: http://www.mayo.edu/hsr/sasmac.html

BITTNER, E. AND GARFINKEL, H. (1967). 'Good' organizational reasons for 'bad' organizational records. *Studies in Ethonomethodology*, Englewood Cliffs, NJ: Prentice Hall, pp. 186–207.

BLUMER, H. (1969). The methodological position of symbolic interactionism. *Symbolic Interactionism: Perspective and Method*, Berkeley, CA: University of California Press.

BOSK, C. L. (1981). *Forgive and Remember: Managing Medical Failure*. Chicago, IL: University of Chicago Press.

BRESLOW, N. E. AND DAY, N. E. (1980). *The Analysis of Case-Control Studies*. Lyon: WHO.

BROOKMEYER, R., LIANG, K. Y. AND LINET, M. (1986). Matched case-control designs and overmatched analyses. *American Journal of Epidemiology* **124**, 693–701.

CHASE, G. R. (1966). On the efficiency of matched pairs in Bernoulli trials. *Biometrika* **55**, 365–369.

COCHRAN, W. G. (1965). The planning of observational studies of human populations (with Discussion). *Journal of the Royal Statistical Society* A **128**, 134–155.

COCHRAN, W. G. (1968). The effectiveness of adustment by subclassificaton in removing bias in observational studies. *Biometrics* **24**, 205–213.

DEHEJIA, R. H. AND WAHBA, S. (1999). Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**, 1053–1062.

EMERSON, R. M. (1981). Observational field work. *Annual Review of Sociology* **7**, 351–378.

ESTROFF, S. E. (1985). *Making It Crazy: An Ethnography of Psychiatric Clients in an American Community*. Berkeley, CA: University of California Press.

GEERTZ, C. (1973). Thick description: toward an interpretative theory of culture. *The Interpretation of Cultures*, Basic Books. pp. 3–30.

GEERTZ, C. (1983). *Local Knowledge*, Basic Books.

GOFFMAN, E. (1974, 1986). *Frame Analysis: An Essay on the Organization of Experience*. Boston, MA: Northeastern University Press.

GU, X. S. AND ROSENBAUM, P. R. (1993). Comparison of multivariate matching methods: structures, distances and

algorithms. *Journal of Computational and Graphical Statistics* **2**, 405–420.

HELPER, S. (2000). Economists and field research. *American Economic Review* **90**, 228–232.

JOFFE, M. M. AND ROSENBAUM, P. R. (1999). Propensity scores. *American Journal of Epidemiology* **150**, 327–333.

KATZ, J. (1999). *How Emotions Work*. Chicago, IL: University of Chicago Press.

KOCH, S. (1999). A theory of definition: implications for psychology, science and the humanities. In Koch, S. (ed.), *Psychology in Human Context: Essays in Dissidence and Reconstruction*, Chicago, IL: University of Chicago Press, pp. 147–191.

LANGHOLZ, B. AND BORGAN, O. (1995). Counter-matching: a stratified nested case-control sampling method. *Biometrika* **82**, 69–79.

MING, K. AND ROSENBAUM, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* **56**, 118–124.

MOSTELLER, F. (1996). Editorial: the promise of risk-based allocation trials in assessing new treatments. *American Journal of Public Health* **86**, 622–623.

POLYA, G. (1954). *Mathematics and Plausible Reasoning, II: Patterns of Plausible Inference*. Princeton, NJ: Princeton University Press.

POPPER, K. R. (1968). *The Logic of Scientific Discovery*. New York: Harper and Row.

QUINE, W. (1992). *Pursuit of Truth*. Cambridge, MA: Harvard University Press.

ROBINS, J. (1989). The control of confounding by intermediate variables. *Statistics in Medicine* **8**, 679–701.

ROSENBAUM, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society* Series A**, 147**, 656–666.

ROSENBAUM, P. R. (1986). Dropping out of high school in the United States: an observational study. *Journal of Educational Statistics* **11**, 207–224.

ROSENBAUM, P. R. (1989). Optimal matching in observational studies. *Journal of the American Statistical Association* **84**, 1024–32.

ROSENBAUM, P. R. (1991a). Discussing hidden bias in observational studies. *Annals of Internal Medicine* **115**, 901–905.

ROSENBAUM, P. R. (1991b). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society* B **53**, 597–610.

ROSENBAUM, P. R. (1995). *Observational Studies*. New York: Springer.

ROSENBAUM, P. R. (1999). Choice as an alternative to control in observational studies (with Discussion). *Statistical Science* **14**, 259–304.

ROSENBAUM, P. AND RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

ROSENBAUM, P. AND RUBIN, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79**, 516–524.

ROSENBAUM, P. AND RUBIN, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* **39**, 33–38.

RUBIN, D. (1973). Matching to remove bias in observational studies. *Biometrics* **29**, 159–183. Correction: 1974, **30**, 728.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* **74**, 318–328.

RYLE, G. (1971). *Collected Papers*, Volume 2. London: Hutchinson.

SHWEDER, R. A. (1996). Quanta and qualia: what is the 'object' of ethnographic method? In Jessor, R., Colby, A. and Shweder, R. (eds), *Ethnography and Human Development*, Chicago, IL: University of Chicago Press, pp. 175–182.

SILBER, J. H., SCHWARTZ, J. S., KRAKAUER, H. AND WILLIAMS, S. V. (1992). Hospital and patient characteristics associated with death after surgery: a study of adverse occurrence and failure to rescue. *Medical Care* **30**, 615–629.

SILBER, J. H., ROSENBAUM, P. R. AND ROSS, R. N. (1995a). Comparing the contributions of groups of predictors: which outcomes vary with hospital rather than patient characteristics? *Journal of the American Statistical Association* **90**, 7–18.

SILBER, J. H., ROSENBAUM, P. R., SCHWARTZ, J. S., ROSS, R. N. AND WILLIAMS, S. V. (1995b). Evaluation of the complication rate as a measure of quality of care in coronary bypass graft surgery. *Journal of the American Medical Association* **274**, 317–323.

SILBER, J. H., ROSENBAUM, P. R., WILLIAMS, S. V., ROSS, R. N. AND SCHWARTZ, J. S. (1997a). The relationship between choice of outcome measure and hospital rank in general surgical procedures: implications for quality assessment. *International Journal for Quality in Health Care* **9**, 193–200.

SILBER, J. H. AND ROSENBAUM, P. R. (1997b). A spurious correlation between hospital mortality and complication rates. *Medical Care* **35**, OS77–OS92.

SILBER, J. H., ROSENBAUM, P. R., KOZIOL, L. F., SUTARIA, N., MARSH, R. R. AND EVEN-SHOSHAN, O. (1999). Conditional length of stay. *Health Services Research* **34**, 349–363.

SMITH, A., KARK, J., CASSEL, J. AND SPEARS, G. (1977). Analysis of prospective epidemiolgic studies by minimum distance case-control matching. *American Journal of Epidemiology* **105**, 567–574.

SMITH, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology* **27**, 325–353.

THOMPSON, W. D., KELSEY, J. L. AND WALTER, S. D. (1982). Cost and efficiency in the choice of matched and unmatched case-control studies. *American Journal of Epidemiology* **116**, 840–851.

WITTGENSTEIN, L. (1994). Anthony Kenny, (ed.), *The Wittgenstein Reader*. Oxford: Blackwell.

[*Received March 15, 2000; revised August 14, 2000; accepted for publication August 31, 2000*]