Published in final edited form as:
J Diabetes Complications. 2013 ; 27(6): . doi:10.1016/j.jdiacomp.2013.07.007.

Observational Research Opportunities and Limitations

Edward J. Boyko, MD, MPH
Epidemiologic Research and Information Center, VA Puget Sound Health Care System, Seattle, WA USA. University of Washington School of Medicine, Seattle, WA

Abstract

Medical research continues to progress in its ability to identify treatments and characteristics associated with benefits and adverse outcomes. The principle engine for the evaluation of treatment efficacy is the randomized controlled trial (RCT). Due to the cost and other considerations, RCTs cannot address all clinically important decisions. Observational research often is used to address issues not addressed or not addressable by RCTs. This article provides an overview of the benefits and limitations of observational research to serve as a guide to the interpretation of this category of research designs in diabetes investigations. The potential for bias is higher in observational research but there are design and analysis features that can address these concerns although not completely eliminate them. Pharmacoepidemiologic research may provide important information regarding relative safety and effectiveness of diabetes pharmaceuticals. Such research must effectively address the important issue of confounding by indication in order to produce clinically meaningful results. Other methods such as instrumental variable analysis are being employed to enable stronger causal inference but these methods also require fulfillment of several key assumptions that may or may not be realistic. Nearly all clinical decisions involve probabilistic reasoning and confronting uncertainly, so a realistic goal for observational research may not be the high standard set by RCTs but instead the level of certainty needed to influence a diagnostic or treatment decision.

A major focus of medical research is the identification of causes of health outcomes, good and bad. The current gold standard method to accomplish this aim is the randomized controlled trial (RCT) (Meldrum, 2000). The performance of a RCT requires strict specification of study conditions related to all aspects of its conduct, such as participant selection, treatment and control assignment arms, inclusion/exclusion criteria, randomization method, outcome measurement, and many other considerations. Such trials are difficult to mount due to the expense in terms of both time and money, and often lead to results that may be difficult to apply to a real-world setting due to either the rigor or complexity of the intervention or the selection process for participants that yields a population dissimilar from that seen in general clinical practice. A randomized controlled trial focuses on an assessment of the validity of its results at the expense of generalizability. For example, the Diabetes Prevention Program screened 158,177 subjects to yield 3,819 subjects who were eventually randomized to one of the four original arms (Rubin et al., 2002). Other limitations of RCTs include a focus on treatment effects and not the ability to detect rarer adverse reactions; restrictions on diabetes duration at the time of trial entry, thereby yielding results that may not apply to persons with a different diabetes duration at the initiation of treatment; and high costs that limits the number of therapeutic comparisons. Regarding this last point,

Corresponding Author Address: 1100 Olive Way, Suite 1400, Seattle WA 98101, P: 206-277-4618 F: 206-764-2563 eboyko@uw.edu.

assessment of a new treatment for hyperglycemia requires comparison to existing accepted treatments, but the control population usually is restricted to fewer treatments than in current use, thereby limiting the ability to compare the new treatment to all existing treatments.

Given these considerations, observational research is often used to address important clinical questions in the absence of randomized clinical trial data, but may also make important potential contributions even when RCTs have been conducted. Examples include monitoring for long-term adverse events that did not appear during the time interval over which the RCT was conducted, or to assess whether the trial findings apply to a different population excluded from the trial due to younger or older age, gender, presence of comorbid conditions, or other factors. Observational research often also addresses other questions not suitable for randomized clinical trials, such as an exposure known to be harmful or in other ways unacceptable to participants or whose administration is inconsistent with ethical principles. Also, observational research can address other exposures that are not potentially under the control of the investigator, such as, for example, eye color, blood type, presence of a specific genetic marker, or elevations of blood pressure or plasma glucose concentration. Observational research may also provide preliminary data to justify the performance of a clinical trial, which might not have received sufficient funding support without the existence of such results.

This paper will review observational research methods applied to addressing questions of causation in diabetes research, with a particular focus on pharmacoepidemiology as an area of research where many important questions may be addressed regarding the relative merits of multiple pharmaceuticals for a given condition. There have been an increasing number of observational studies of the association between diabetes treatments and hard outcomes, such as death or CVD events. The increase in such studies likely has been facilitated by the availability of big data in general and specifically large pharmaceutical databases created by national health plans, large health care systems, or mail-order pharmacy providers (Sobek et al., 2011). In addition, the ongoing development of diabetes pharmacotherapies approved based on ability to achieve an improvement in glycemic control but without data on hard outcomes may also provide the impetus to use such large databases for research on comparative safety and efficacy.

Observational Research Study Designs

Cohort and Case-Control Studies

The two most popular designs for investigating causal hypotheses are the cohort and case-control studies. Features are shown in Table 1. The major difference between the two is that the cohort study begins with identification of the exposure status, whereas the case-control study begins with the identification of the outcome. A cohort study can be prospective, where exposed and non-exposed subjects are followed for the development of the outcome, or retrospective, where collected data can be used to identify both the exposure status at some past time point and the subsequent development of the outcome. A case-control study, on the other hand, can only look back in time for occurrence of the exposure. There are of course exceptions to these general statements. It is possible in some case-control studies to measure the exposure after the outcome in time if the exposure is invariant and if it is not related to a greater loss to follow-up among persons with the outcome due to mortality or other reasons. Examples of such exposures include genetic markers or an unchanging characteristic of adults such as femur length, eye color, or red blood cell type. Variations in these study designs include the case-cohort and case-only studies, which are described in detail elsewhere, and which a description of which will not be provided here (DiPietro, 2010). Also, the relative merits of these study designs will not be discussed here but are covered in standard epidemiology texts.

Weaker Observational Research Designs

Other research designs are often used in studies reported in the medical literature. These include cross-sectional, case-series, and case-reports. The cross-sectional study has limited value in assessing a potential causal relationship since it may not be possible to determine whether the potential exposure preceded the outcome, except when the exposure does not vary over one's life history, such as in the case of a genotype, ABO blood group, or eye color. Case-series and case reports are even more limited since it is not possible to assess if the outcome occurred more frequently among the persons included compared to a control population. Case reports do though have potential value in pharmaceutical safety research by generating potential signals that signify unexpected adverse events. Such monitoring is employed in the Food and Drug Administration's Adverse Event Reporting System, and has led to changes in product labeling as well as restriction or outright removal of pharmaceuticals from the market due to safety concerns (Wysowski & Swartz, 2005). Over 2 million case reports of adverse reactions were submitted between 1969–2002, resulting in only about 1% of marketed drugs being withdrawn or restricted. Therefore the noise-to-signal ratio for this method of surveillance is exceedingly high and presents an opportunity for other observational methods to better address this issue.

Observation Research for Causal Inference

Causal associations will always involve correlation, but the presence of a correlation does not imply causation. The challenge of observational research is to assess whether a correlation is present and then determine whether it may be due to a causal association. A list of criteria was developed by Dr. Austin Bradford Hill decades ago that is still referred to frequently today (Hill, 1965), although reexamination of these criteria more recently has led to the conclusion that only one of the nine original features is really necessary for a causal relationship in a observational study (Phillips & Goodman, 2004; Rothman & Greenland, 2005). The magnitude of the observed association, another Hill criterion, often figures into determinations about the presence of bias, with those of greater magnitude considered less likely to be due to bias and more likely due to a causal process (Grimes & Schulz, 2002).

Examination of the features of an RCT provide some insight into the limitations of observational research in assessing causal associations. The randomization process provides the opportunity for equal distribution of risk factors for the outcome among persons assigned to the treatment and control. Thus any difference in the outcome between these two groups will not likely be due to unequal distribution of risk factors by treatment assignment. The use of randomization provides a way to approach the problem of not having complete knowledge about predictors of all clinically important outcomes. If we did have such knowledge then groups with exactly equal risks of the outcome could be assembled by the investigator. As we do not have such knowledge, the process of randomization utilizes chance to distribute both known and more importantly unknown risk factors for the outcome, and is most likely to achieve this aim with larger sample size (Efird, 2011). Randomization, though, does not guarantee that the treatment and control group will have the same risk of the outcome. Accidents of randomization have occurred for known risk factors for outcomes as in the UGDP, where older subjects with a higher prevalence of cardiovascular disease risk factors were disproportionately assigned to the tolbutamide treatment arm (Leibel, 1971). Such accidents also must occur for the unknown risk factors, although these would not be apparent to the investigator.

Bias in Observational Research

Confounding Bias

Observational research does not have the benefit of randomization to allocate by chance risk factors for an outcome of interest. Exposures to risk factors occur due to self-selection, medical provider prescription, in association with occupation, and for other reasons. When an exposure of interest is strongly associated with another exposure that is also related to the outcome, confounding bias is present, but methods exist to obtain an unbiased estimated of the exposure-disease association as long as the confounding factor is identified and measured accurately.

A cross-sectional study of a genetic marker (Gm haplotype $Gm^{3;5,13,14}$) and diabetes prevalence provides an example of confounding bias. Subjects included members of the Pima and Papago tribes of the Gila River Indian Community in Southern Arizona who underwent a medical history and examination every two years including assessment of diabetes status through oral glucose tolerance testing (Knowler, Williams, Pettitt & Steinberg, 1988). Subjects were further characterized by degree of Indian heritage measured in eighths and referred to as “quantum.” A total of 4,640 subjects of either 0/8, 4/8, and 8/8 quantum were included in this analysis. There were 1,336 persons with and 3,304 persons without diabetes available for analysis, yielding a crude (unadjusted) overall odds ratio of 0.24 for the association between haplotype $Gm^{3;5,13,14}$ and diabetes prevalence (Figure 1, Panel A). This result supports a lower prevalence of diabetes in association with haplotype $Gm^{3;5,13,14}$, but the unadjusted result represents a substantial overestimate due to confounding by Quantum. In Figure 1 panel B, subjects were divided by the three Quantum categories found in the sample, and within each of these the odds ratio is closer to 1.0 and therefore of smaller magnitude than the crude result. Note that collapsing the three tables in Panel B by summing the cells yields the single overall table shown in Panel A. Adjustment for these Quantum categories yields an odds ratio of 0.59, which is of smaller magnitude than the result seen in the unadjusted analysis (Figure 1, Panel B). Although the odds ratios vary across Quantum categories, a test for heterogeneity across these strata was non-significant (p=0.295). Therefore the null hypothesis that the odds ratios differed across Quantum strata could not be rejected.

Examination of the frequency of haplotype $Gm^{3;5,13,14}$ and diabetes prevalence across Indian heritage Quantum reveals the reason for the overestimation of the association in the unadjusted analysis. Diabetes occurred more frequently while the haplotype $Gm^{3;5,13,14}$ occurred less frequently among subjects with greater Indian heritage (Figure 1, Panel C). Adjustment for the imbalance in Quantum by haplotype $Gm^{3;5,13,14}$ in this specific example and in general any accurately measured confounding factor yields a less biased odds ratio that is closer to the true magnitude of the association between this haplotype and diabetes prevalence.

Another more recent example of confounding can be seen in a case-cohort European study of the association between artificially sweetened soft drinks and the risk of developing type 2 diabetes (2013). The unadjusted hazard ratio for the daily consumption of ≥ 250 g of this beverage type was 1.84 (95% CI 1.52 to 2.23) representing a statistically significant elevation in risk. After adjustment for daily energy intake and BMI, the hazard ratio diminished to 1.13 (95% CI 0.85 to 1.52) and was no longer statistically significant (p=0.24). The investigators concluded that consumption of artificially sweetened soft drinks was not associated with type 2 diabetes risk in their population.

Multiple methods exist to remove the bias from recognized, accurately measured confounding factors, but unfortunately there is no widely accepted option for handling

unmeasured confounding factors and adjusting for this bias. In this regard observational research is unable to match the ability of a RCT to account for this potential bias. Methods have been developed to better assess whether associations represent causal pathways that will be described later in this paper.

Information Bias

Observational research can be susceptible to other types of bias. Information bias refers to inaccurate assessment of the outcome, the exposure, or potential confounding variables. An example includes measurement of nutritional intake, which is often assessed by research subjects completing a food frequency survey or 24-hour dietary recall. Even if subjects report these intakes correctly, the likelihood is low that this will reflect long-term dietary intake exactly. Attempts have been made to reduce the error of these measurements through biomarker calibration that in one study was based on a urinary nitrogen protocol to estimate daily protein consumption over a 24-hour period (Tinker et al., 2011). This analysis revealed a slight increase in risk of incident diabetes in association with a 20% higher protein intake in grams (Hazard Ratio 1.05, 95% CI 1.03–1.07). Recalibrated results based on the results of the urinary nitrogen protocol yielded a substantially higher diabetes hazard ratio of 1.82 (95% CI 1.56–2.12) that after adjustment for BMI was reduced to 1.16 (95% CI 1.05–2.28). In this example, reduction of measurement error yielded a difference of greater magnitude than see in the analysis based on dietary self-reports only without objective validation, although theoretically more accurate measurements may yield smaller differences, depending on the type and magnitude of measurement error.

Selection Bias

Selection bias may produce factitious exposure-disease associations if the study population fails to mirror the target population of interest. For example, selection of control subjects from among hospitalized patients as might be the case in a study based on administrative data may not accurately depict smoking prevalence among controls, given that smoking is related to multiple diseases that would increase the risk for hospitalization. Effective observational research must recognize the potential for bias and attempt to minimize it both in the design and analysis, as well as accurately describing limitations of these data and the implications for study validity in reports of results.

Agreement and Discrepancies between Observational and Clinical Trial Research

One way to assess whether the potential biases of observational studies result in failure to detect true associations is by comparison of observational versus RCT results on the same questions. Since observational studies of treatments often precede definitive clinical trials, several authors have assessed agreement between similar hypotheses tested using the gold standard compared to observational designs, concluding that agreement between the two is high. A comparison of 136 reports published between 1985 to 1998 on 19 different treatments found excellent agreement, with the combined magnitude of the effect in observational studies lying within the 95% confidence intervals of the combined magnitude of the effect in RCTs for 17 of the 19 hypotheses tested (Benson & Hartz, 2000). Another comparison focused on comparing the results of meta-analyses of observation and clinical trial research on five clinical questions that were identified through a search of five major medical journals from 1991 to 1995 (Concato, Shah & Horwitz, 2000). These investigators concluded that average results of these studies were “remarkably similar.”

In contrast, other research has demonstrated discrepancies between RCT and observational designs. The Women’s Health Initiative (WHI) was a RCT of dietary and menopausal

hormone interventions to assess these effects on mortality, cardiovascular disease, and cancer risk (Prentice et al., 2005). Perhaps unique to this study was the establishment of a concurrent observational study accompanying the randomized clinical trial, thereby permitting direct comparison of reported associations by type of research design within the same study framework. In the trial/observational study of estrogen plus progestin for menopausal hormone replacement, marked differences were seen between the treatment and control groups by participation in the RCT or observational study (Table 2). In the RCT, no important differences were seen by treatment assignment for race, educational level, BMI, or current smoking status. This was not true by estrogen-progestin exposure in the observational study, where exposed women were more likely to be White, having completed a college degree or higher, and less likely to be current smokers or obese. Outcomes occurred more frequently in the estrogen-progestin arm of the RCT, but less frequently in the corresponding arm of the observational study, except for venous thromboembolism (Table 2). Hazard ratios for these comparisons adjusted for imbalances in baseline potential confounding factors show a harmful effect of estrogen-progestin use that is statistically significantly elevated in 2 of 3 outcomes and a discordance with the observational results due to null, somewhat protective hazard ratios or in the case of venous thromboembolism, an elevated hazard ratio of considerably smaller magnitude than in the clinical trial. Although good agreement between clinical trials and observational research occurs often, the example of the WHI prevents having complete confidence in the results of observational studies.

Achievements of Observational Research

Despite the limitations of observational research design, many well-accepted causal associations in medicine are supported entirely or in part due to this type of investigation. Several examples include the association between hyperglycemia and diabetes complications including retinopathy, nephropathy, peripheral neuropathy, and ischemic heart disease (2013). Other well known examples include hypertension and stroke, smoking and lung cancer, asbestosis and mesothelioma, and LDL and HDL cholesterol concentrations and risk of ischemic heart disease (Churg, 1988; Gordon, Kannel, Castelli & Dawber, 1981; Kannel, Wolf, Verter & McNamara, 1970; Pirie, Peto, Reeves, Green & Beral, 2013). In the case of complications due to hyperglycemia, high LDL-cholesterol concentration, and hypertension, clinical trials to reduce these levels have resulted in reductions in the rate of these outcomes, further supporting a causal association (1991; 1994; 1998; 1998). For many associations that involve an exposure that cannot be controlled by the investigator or should not be modified for ethical reasons, observational research may be the only avenue for direct testing of these associations in humans.

Causal Inference from Observational Research

The results of an observational research study are never interpreted in an information vacuum. Given the potential for bias with this study design, a number of other factors should be considered when weighing the strength of this evidence. First and foremost would be the replication of the finding in other observational research studies. Additional evidence to bolster the potential causal association would be support from the biological understanding of underlying mechanisms, animal experiments confirming that the exposure results in a similar outcome, and trend data in disease incidence following changes in exposure prevalence. For example, in the UK Million Women Study where median age was reported at 55 years, women who quit smoking completely at ages 25–34 or 35–44 years had only 3% and 10% of the excess mortality, respectively, seen among women who were continuing smokers (Pirie, Peto, Reeves, Green & Beral, 2013). Coronary heart disease deaths in the U.S. declined by approximately 50% between 1980 to 2000. One analysis that addressed the reasons for this decline concluded that change in risk factors (reductions in total cholesterol

concentration, systolic blood pressure, smoking, and physical inactivity) accounted for approximately 47% of this decrease (Ford et al., 2007). These trends provide support for a causal association between smoking and lung cancer, and multiple cardiovascular disease risk factors and coronary death risk.

Pharmacoepidemiology

Many questions regarding the use of pharmaceuticals may never be answered through use of RCTs, thereby creating a need to address knowledge gaps using observational research. The specialized field of pharmacoepidemiology directly addresses these needs. The earliest appearance of the term “pharmacoepidemiology” on PubMed.com is in an article written in 1984 (Lawson, 1984). The field of pharmacoepidemiology encompasses the use of observational research to assess pharmaceutical safety and effectiveness. For example, diabetes pharmaceuticals have received FDA approval based on efficacy at lowering glucose and safety, without the need to prove efficacy at preventing long-term complications. The sulfonylurea hypoglycemic agents glyburide and glipizide are in widespread use to manage the hyperglycemia of diabetes, but it is not clear whether one is associated with a greater reduction in hard outcomes such as mortality or diabetes complications, as this has not been tested in a clinical trial. Use of such surrogate endpoints as opposed to the hard outcomes one wishes to prevent has been criticized as an ineffective and potentially harmful approach to medication approval (Fleming & DeMets, 1996; Psaty et al., 1999). Design of clinical trials to address hard as opposed to surrogate endpoints typically requires larger sample size, longer follow-up, and greater costs.

Observational research may also identify adverse effects associated with the use of pharmaceuticals that were not anticipated based on research conducted in support of the drug approval process. The withdrawal of the thiazolidinedione agent troglitazone from the U.S. market in 2000 followed reports on cases of severe liver toxicity during post-marketing surveillance. Similar data on a high number of reported cases of severe myopathies in cerivastatin users led to its withdrawal from the worldwide market in 2001 (Furberg & Pitt, 2001). An observational study using administrative claims databases to assess the relative safety of lipid lowering medications in the U.S. between 2000–2004 reported a much higher risk for hospitalization for treatment of myopathy among cerivastatin users compared to users of other statin and non-stain lipid lowering agents (Cziraky et al., 2006).

Confounding by Indication

As with other observational research designs, there are limitations to pharmacoepidemiology due to biases previous described, but in addition to these is the vexing phenomenon of confounding by indication, also referred to as channeling bias (McMahon & MacDonald, 2000; Petri & Urquhart, 1991). This refers to an observed benefit (or harm) associated with a pharmaceutical due to the indications for treatment with it and not a medication effect. A hypothetical example of how confounding by indication results in outcome differences not due to medication effect is shown in Figure 2, which provides an example of how the choice of a diabetes pharmaceutical may depend on the existence of a condition (higher serum creatinine reflecting lower GFR) associated with higher mortality risk (Fox et al., 2012).

Several approaches exist to the problem of confounding by indication. If there is no association between the indication for the pharmaceutical and the outcome of interest, then no bias will occur, since an association must also be present between both the indication and the outcome to yield a biased result. This same principle applies to all confounding factors (van Stralen, Dekker, Zoccali & Jager, 2010). If the conditions for confounding are fulfilled, then statistical adjustment techniques are available to produce unbiased estimates of effect. Commonly used methods in biomedical research include linear regression analysis for

continuous outcomes, logistic regression for categorical outcomes, and the Cox proportional hazards model for time-to-event outcomes. In addition, propensity scores have risen in popularity over the past decade. An “all fields” search of Pubmed conducted January 15, 2012 using the search term “propensity score” yielded 2,895 hits for the immediate past 5 years, and only 715 hits for the previous 5 years. The propensity score method models the probability of exposure in relation to predictor variables, and therefore estimates the likelihood, in the case of a pharmacoepidemiology study, of a subject receiving a particular pharmaceutical based on his or her characteristics (Rubin, 2010). An additional step is required which uses standard previously mentioned adjustment methods to remove the bias associated with varying likelihood of receiving the pharmaceutical. Despite the rising popularity of this method, it has been demonstrated to be merely equivalent and sometimes inferior to standard multivariate adjustment methods (Shah, Laupacis, Hux & Austin, 2005; Sturmer et al., 2006). Furthermore, propensity scores cannot address the issue of unmeasured confounding (Cummings, 2008). So if the indications for the pharmaceutical cannot be determined from the other measured factors, neither multivariate adjustment or propensity scores will allow for adjustment and removal of bias.

Several design features of observational studies may increase the likelihood of confounding by indication but if recognized may be amenable to correction in the design or analysis phases of a study. Assessing outcomes for pharmaceuticals prescribed for different indications or by a comparison of populations who differ with regard to the presence of medication contraindications may introduce bias into comparisons. An assessment of the mortality risk associated with beta-blocker use compared to other antihypertensive medications should exclude participants in whom beta blockers but not other antihypertensive medications are prescribed for other indications, such as migraine headache or stage fright prophylaxis, as these conditions may be associated with better outcomes and lead to over-optimistic survival benefit. Also, failure to consider medication contraindications may lead to risk of the outcome differing by medication used, as seen in the example in Figure 2 which would lead to a higher frequency of subjects with renal insufficiency in the glipizide treatment group for hyperglycemia. To account for this potential bias, subjects with contraindications for use of any of the pharmaceuticals of interest in the comparison should be eliminated from the study. For example, recent studies of mortality and cardiovascular events among users of sulfonylurea or metformin monotherapy for treatment of diabetes in the Veterans Health Administration system excluded patients with serious medical conditions at baseline that might influence the prescription of diabetes medication (Roumie et al., 2012; Wheeler et al., 2013). For example, some items on the list of exclusions were congestive heart failure, serum creatinine concentration of 1.5 mg/dl or greater, HIV, and other conditions described in this publication. Despite these design features and adjustment methods to correct for factors associated with a particular prescription that may also be associated with a different outcome risk, there will always be some uncertainty about the presence of bias due to residual confounding by indication.

Methods to Improve Causal Inference from Observational Research

Instrumental variables analysis has been promoted as a method to overcome the inability to exclude undetected confounding in observational research. This method involves identification of a factor that strongly predicts treatment (or exposure in an epidemiologic study not involving a pharmaceutical). This factor is referred to as an “instrument,” and it is used in a manner analogous to the intention to treat analysis employed in RCTs (Thomas & Conti, 2004). A Mendelian Randomization study is a type of instrumental variable analysis that uses a genetic marker as the instrument (Thomas & Conti, 2004). Although intriguing in concept, the difficulty is in the application, as this relies on finding an “instrument” that is

(1) causally related to treatment but not unobserved risk factors for the outcome, and (2) influences the outcome only through its effect on treatment (Hernan & Robins, 2006). This method is being explored in pharmacoepidemiologic investigations, with one example being use of physician prescribing preference for types of NSAIDS in the evaluation of the gastrointestinal toxicity of COX-2 inhibitors versus non-COX-2 inhibitor NSAIDS (Brookhart, Wang, Solomon & Schneeweiss, 2006). This analysis reported a protective association with COX-2 inhibitors only in the instrumental variable analysis, leading the authors to conclude that this analysis resulted in a reduction in unmeasured confounding. Examples can also be found in the diabetes epidemiology literature, such as the lack of association between serum uric acid level and type 2 diabetes risk (Pfister et al., 2011), and higher risk associated with lower sex hormone-binding globulin concentration (Ding et al., 2009).

Conclusions

As it will not be possible to assess efficacy of all possible treatment comparisons in all possible groups of interest, or identify adverse (or unexpected beneficial) outcomes requiring longer follow-up or greater sample size using RCTs, observational research stands prepared to step forward to address these knowledge gaps. Much medical knowledge and practice currently rests on a foundation of observational research. Perhaps this is not noticed due to the gloss and novelty of recently completed RCTs. Little research has been conducted comparing results from observational and clinical trial designs, but that which has been completed finds generally good agreement in these findings. With any observational research finding, though, comes less certainly due to the inability to completely exclude the possibility of residual confounding, or in the case of a pharmaceutical, confounding by indication. However, the expectation of absolute certainty is unrealistic and inconsistent with the current practice of medicine, where decisions are made probabilistically, with the threshold for actions such as further testing or treatment varying widely depending on the comparative costs and benefits of true and false positive and negative decisions (Boland & Lehmann, 2010; Pauker & Kassirer, 1980; Plasencia, Alderman, Baron, Rolfs & Boyko, 1992). Observational research definitely has had and will continue to have an important role in providing the information needed to improve medical decision-making. There is always room for improvement and the hope that the future will bring better methods to further reduce the uncertainty surrounding the validity of its results.

Acknowledgments

Grant Support: VA Epidemiologic Research and Information Center; the Diabetes Research Center at the University of Washington (DK-017047)

Thanks for James S. Floyd MD for his careful review of this manuscript. The work was supported by the VA Epidemiologic Research and Information Center; the Diabetes Research Center at the University of Washington (DK-017047); and VA Puget Sound Health Care System.

References

Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension. Final results of the Systolic Hypertension in the Elderly Program (SHEP). SHEP Cooperative Research Group. JAMA. 1991; 265:3255–3264. [PubMed: 2046107]

1994 Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). Lancet. 344:1383–1389. [PubMed: 7968073]

Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34). UK Prospective Diabetes Study (UKPDS) Group. Lancet. 1998; 352:854–865. [PubMed: 9742977]

Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). UK Prospective Diabetes Study (UKPDS) Group. Lancet. 1998; 352:837–853. [PubMed: 9742976]

Consumption of sweet beverages and type 2 diabetes incidence in European adults: results from EPIC-InterAct. Diabetologia. 2013; 56:1520–1530. [PubMed: 23620057]

Diagnosis and classification of diabetes mellitus. Diabetes Care. 2013; 36(Suppl 1):S67–74. [PubMed: 23264425]

Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med. 2000; 342:1878–1886. [PubMed: 10861324]

Boland MV, Lehmann HP. A new method for determining physician decision thresholds using empiric, uncertain recommendations. BMC Med Inform Decis Mak. 2010; 10:20. [PubMed: 20377882]

Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. Epidemiology. 2006; 17:268–275. [PubMed: 16617275]

Churg A. Chrysotile, tremolite, and malignant mesothelioma in man. Chest. 1988; 93:621–628. [PubMed: 2830081]

Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000; 342:1887–1892. [PubMed: 10861325]

Cummings P. Propensity scores. Arch Pediatr Adolesc Med. 2008; 162:734–737. [PubMed: 18678805]

Cziraky MJ, Willey VJ, McKenney JM, Kamat SA, Fisher MD, Guyton JR, et al. Statin safety: an assessment using an administrative claims database. Am J Cardiol. 2006; 97:61C–68C. [PubMed: 16377285]

Ding EL, Song Y, Manson JE, Hunter DJ, Lee CC, Rifai N, et al. Sex hormone-binding globulin and risk of type 2 diabetes in women and men. N Engl J Med. 2009; 361:1152–1163. [PubMed: 19657112]

DiPietro NA. Methods in epidemiology: observational study designs. Pharmacotherapy. 2010; 30:973–984. [PubMed: 20874034]

Efird J. Blocked randomization with randomly selected block sizes. Int J Environ Res Public Health. 2011; 8:15–20. [PubMed: 21318011]

Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med. 1996; 125:605–613. [PubMed: 8815760]

Ford ES, Ajani UA, Croft JB, Critchley JA, Labarthe DR, Kottke TE, et al. Explaining the decrease in U.S. deaths from coronary disease, 1980–2000. N Engl J Med. 2007; 356:2388–2398. [PubMed: 17554120]

Fox CS, Matsushita K, Woodward M, Bilo HJ, Chalmers J, Heerspink HJ, et al. Associations of kidney disease measures with mortality and end-stage renal disease in individuals with and without diabetes: a meta-analysis. Lancet. 2012; 380:1662–1673. [PubMed: 23013602]

Furberg CD, Pitt B. Withdrawal of cerivastatin from the world market. Curr Control Trials Cardiovasc Med. 2001; 2:205–207. [PubMed: 11806796]

Gordon T, Kannel WB, Castelli WP, Dawber TR. Lipoproteins, cardiovascular disease, and death. The Framingham study. Arch Intern Med. 1981; 141:1128–1131. [PubMed: 7259370]

Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet. 2002; 359:248–252. [PubMed: 11812579]

Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology. 2006; 17:360–372. [PubMed: 16755261]

Hill AB. The Environment and Disease: Association or Causation? Proceedings of the Royal Society of Medicine. 1965; 58:295–300. [PubMed: 14283879]

Kannel WB, Wolf PA, Verter J, McNamara PM. Epidemiologic assessment of the role of blood pressure in stroke. The Framingham study. JAMA. 1970; 214:301–310. [PubMed: 5469068]

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. American journal of human genetics. 1988; 43:520–526. [PubMed: 3177389]

Lawson DH. Pharmacoepidemiology: a new discipline. Br Med J (Clin Res Ed). 1984; 289:940–941.

Leibel B. An analysis of the University Group Diabetes Study Program: data results and conslusions. Can Med Assoc J. 1971; 105:292–294. [PubMed: 5563349]

McMahon AD, MacDonald TM. Design issues for drug epidemiology. Br J Clin Pharmacol. 2000; 50:419–425. [PubMed: 11069436]

Meldrum ML. A brief history of the randomized controlled trial. From oranges and lemons to the gold standard. Hematol Oncol Clin North Am. 2000; 14:745–760. vii. [PubMed: 10949771]

Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med. 1980; 302:1109–1117. [PubMed: 7366635]

Petri H, Urquhart J. Channeling bias in the interpretation of drug effects. Stat Med. 1991; 10:577–581. [PubMed: 2057656]

Pfister R, Barnes D, Luben R, Forouhi NG, Bochud M, Khaw KT, et al. No evidence for a causal link between uric acid and type 2 diabetes: a Mendelian randomisation approach. Diabetologia. 2011; 54:2561–2569. [PubMed: 21717115]

Phillips CV, Goodman KJ. The missed lessons of Sir Austin Bradford Hill. Epidemiol Perspect Innov. 2004; 1:3. [PubMed: 15507128]

Pirie K, Peto R, Reeves GK, Green J, Beral V. The 21st century hazards of smoking and benefits of stopping: a prospective study of one million women in the UK. Lancet. 2013; 381:133–141. [PubMed: 23107252]

Plasencia CM, Alderman BW, Baron AE, Rolfs RT, Boyko EJ. A method to describe physician decision thresholds and its application in examining the diagnosis of coronary artery disease based on exercise treadmill testing. Med Decis Making. 1992; 12:204–212. [PubMed: 1513211]

Prentice RL, Langer R, Stefanick ML, Howard BV, Pettinger M, Anderson G, et al. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. Am J Epidemiol. 2005; 162:404–414. [PubMed: 16033876]

Psaty BM, Weiss NS, Furberg CD, Koepsell TD, Siscovick DS, Rosendaal FR, et al. Surrogate end points, health outcomes, and the drug-approval process for the treatment of risk factors for cardiovascular disease. JAMA. 1999; 282:786–790. [PubMed: 10463718]

Rothman KJ, Greenland S. Causation and causal inference in epidemiology. Am J Public Health. 2005; 95(Suppl 1):S144–150. [PubMed: 16030331]

Roumie CL, Hung AM, Greevy RA, Grijalva CG, Liu X, Murff HJ, et al. Comparative effectiveness of sulfonylurea and metformin monotherapy on cardiovascular events in type 2 diabetes mellitus: a cohort study. Ann Intern Med. 2012; 157:601–610. [PubMed: 23128859]

Rubin DB. Propensity score methods. Am J Ophthalmol. 2010; 149:7–9. [PubMed: 20103037]

Rubin RR, Fujimoto WY, Marrero DG, Brenneman T, Charleston JB, Edelstein SL, et al. The Diabetes Prevention Program: recruitment methods and results. Control Clin Trials. 2002; 23:157–171. [PubMed: 11943442]

Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol. 2005; 58:550–559. [PubMed: 15878468]

Sobek M, Cleveland L, Flood S, Hall PK, King ML, Ruggles S, et al. Big Data: Large-Scale Historical Infrastructure from the Minnesota Population Center. Hist Methods. 2011; 44:61–68. [PubMed: 21949459]

Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol. 2006; 59:437–447. [PubMed: 16632131]

Thomas DC, Conti DV. Commentary: the concept of 'Mendelian Randomization'. Int J Epidemiol. 2004; 33:21–25. [PubMed: 15075141]

Tinker LF, Sarto GE, Howard BV, Huang Y, Neuhouser ML, Mossavar-Rahmani Y, et al. Biomarker-calibrated dietary energy and protein intake associations with diabetes risk among postmenopausal women from the Women's Health Initiative. Am J Clin Nutr. 2011; 94:1600–1606. [PubMed: 22071707]

van Stralen KJ, Dekker FW, Zoccali C, Jager KJ. Confounding. Nephron Clin Pract. 2010; 116:c143–147. [PubMed: 20516714]

Wheeler S, Moore K, Forsberg CW, Riley K, Floyd JS, Smith NL, et al. Mortality among veterans with type 2 diabetes initiating metformin, sulfonylurea or rosiglitazone monotherapy. Diabetologia. 2013

Wysowski DK, Swartz L. Adverse drug event surveillance and drug withdrawals in the United States, 1969–2002: the importance of reporting suspected reactions. Arch Intern Med. 2005; 165:1363–1369. [PubMed: 15983284]

A

	Diabetes Present	Diabetes Absent
$Gm^{3;5,13,14}$ Present	22	213
$Gm^{3;5,13,14}$ Absent	1314	3091
	OR = 0.24	

B

	0/8 Quantum		4/8 Quantum		8/8 Quantum	
	Diabetes Present	Diabetes Absent	Diabetes Present	Diabetes Absent	Diabetes Present	Diabetes Absent
$Gm^{3;5,13,14}$ Present	3	18	7	138	12	57
$Gm^{3;5,13,14}$ Absent	2	9	8	191	1304	2891
	OR = 0.75		OR = 1.21		OR = 0.47	
			$OR_{adjusted}$ = 0.59			

C

	0/8 Quantum	4/8 Quantum	8/8 Quantum
$Gm^{3;5,13,14}$ Present	21/32 = 0.66	145/344 = 0.42	69/4264 = 0.02
Diabetes Present	5/32 = 0.16	15/344 = 0.04	1316/4264 = 0.31

Figure 1.

Cross-sectional study of Native Americans of the Pima and Papago Indian tribes in Southern Arizona on the associations between the GM haplotype $Gm^{3;5,13,14}$, native quantum, and diabetes mellitus prevalence. Panel A displays all participants combined with Native quantum of either 0/8, 4/8 or 8/8 by presence of diabetes mellitus in relation to $Gm^{3;5,13,14}$ presence or absence. The overall (crude) odds ratio for the association is shown. Panel B displays all participants from Panel A stratified by Native quantum, demonstrating confounding by Native quantum as judged by the discordance between the crude and stratified or Quantum-adjusted results. Panel C demonstrates that Quantum meets the criterion as a confounding variable due to its negative association with $Gm^{3;5,13,14}$ and positive association with diabetes prevalence.

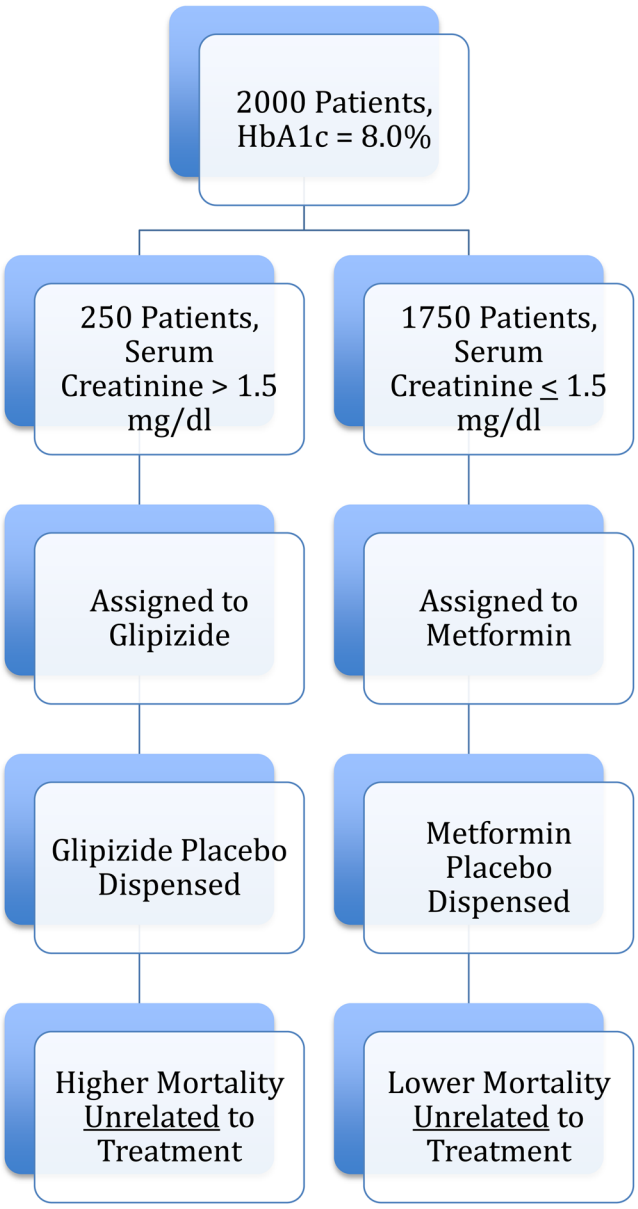

Figure 2.
A hypothetical population of 2000 identical persons with type 2 diabetes differing only by renal function as measured by serum creatinine and assigned to either metformin or glipizide based on the serum creatinine level. The active treatment, though, is never dispensed, and instead substituted with a identical placebo. An expected difference in mortality is seen between the two groups given the association between poorer renal function and mortality in the glipizide group. This difference cannot be explained by the effect of the active pharmaceutical (since there was none) and therefore represents an example of confounding by indication.

Table 1

Observational Study Designs for Assessment of Causal Relationships

	Subjects Selection Based on	Temporal Sequence	Strengths	Weaknesses	Best Application
Cohort	Exposure status	Disease assessed following exposure	Generally less measurement error; provides estimate of incidence; can examine multiple outcomes	May require larger samples sizes and long follow-up	Exposure is rare; multiple outcomes of interest; common outcome(s)
Case-control	Disease status	Exposure assessed prior to disease	Multiple exposures can be examined; smaller sample size needed	Can only examine a single disease of interest; greater potential for bias in measuring exposure	Disease is rare; single disease of interest; common exposure
Cross-sectional	Neither exposure nor disease	Both exposure and disease assessed at the same time point	Straightforward subject selection	Lack of information on exposure timing and disease onset	Quick execution

Table 2

Comparison of baseline characteristics and outcomes in the randomized controlled trial and observational study of estrogen-progestin treatment in the Women's Health Initiative (1994–2002).

	Clinical Trial		Observational Study	
	Placebo Control	**Estrogen-Progestin**	**Control**	**Estrogen-Progestin**
Baseline Characteristics				
White Race	83.9%	84.0%	82.3%	89.2%
Obese	34.0%	34.2%	27.3%	15.7%
College Degree or Higher	35.3%	34.5%	42.9%	53.4%
Current Smoker	10.5%	10.4%	7.0%	4.7%
Outcomes *				
Coronary Heart Disease	0.33	0.40	0.28	0.20
Stroke	0.24	0.32	0.22	0.17
Venous Thromboembolism	0.17	0.35	0.16	0.17
Outcomes	Adjusted HR †, 95% CI		Adjusted HR, 95% CI	
Coronary Heart Disease	1.27 (1.00–1.61)		0.87 (0.72–1.05)	
Stroke	1.21 (0.93–1.59)		0.86 (0.70–1.07)	
Venous Thromboembolism	2.13 (1.59–2.85)		1.31 (1.07–1.61)	

* Age-adjusted annualized incidence (%)

† Adjusted for age, race, education, BMI, smoking, age at menopause, and physical functioning