Effective Health Care Program
Research Reports

Number 32

The Incident User Design in Comparative Effectiveness Research

Eric S. Johnson, Ph.D.
Barbara A. Bartman, M.D., M.P.H.
Becky A. Briesacher, Ph.D.
Neil S. Fleming, Ph.D.
Tobias Gerhard, Ph.D.
Cynthia J. Kornegay, Ph.D.
Parivash Nourjah, Ph.D.
Brian Sauer, Ph.D.
Glen T. Schumock, Pharm.D., M.B.A.
Art Sedrakyan, M.D., Ph.D.
Til Stürmer, M.D., M.P.H.
Suzanne L. West, Ph.D., M.P.H.
Sebastian Schneeweiss, M.D., Sc.D.

Research from the Developing Evidence to Inform Decisions about Effectiveness (DEcIDE) Network

DEPARTMENT OF HEALTH & HUMAN SERVICES • USA

AHRQ
Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

May 2012

The DEcIDE (Developing Evidence to Inform Decisions about Effectiveness) network is part of AHRQ's Effective Health Care Program. It is a collaborative network of research centers that support the rapid development of new scientific information and analytic tools. The DEcIDE network assists health care providers, patients, and policymakers seeking unbiased information about the outcomes, clinical effectiveness, safety, and appropriateness of health care items and services, particularly prescription medications and medical devices.

This report is based on research conducted under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract Nos. HHSA 290-2007-10057I-11, HHSA 290-2005-00161-6, and HHSA 290-2005-00161-10). The AHRQ Task Order Officer for this project was Parivash Nourjah, Ph.D.

The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or the U.S. Department of Health and Human Services.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

None of the authors has a financial interest in any of the products discussed in this report.

Suggested citation:
Johnson ES, Bartman BA, Briesacher BA, Fleming NS, Gerhard T, Kornegay CJ, Nourjah P, Sauer B, Schumock GT, Sedrakyan A, Stürmer T, West SL, Schneeweiss S. The Incident User Design in Comparative Effectiveness Research. Effective Health Care Program Research Report No. 32. (Prepared under Contract No. HHSA290200500161). AHRQ Publication No. 11(12)-EHC054-EF. Rockville, MD: Agency for Healthcare Research and Quality. May 2012. http://effectivehealthcare.ahrq.gov/reports/final.cfm.

Effective Health Care Program Research Report Number 32

The Incident User Design in Comparative Effectiveness Research

Structured Abstract

Background. When conducting comparative effectiveness research with cohort studies or registries, some investigators restrict enrollment to patients who were incident users of an intervention; other investigators enroll all patients who used an intervention. An incident user design follows patients from the day that they started an intervention, which may reduce biases, such as confounding. But an incident user design also reduces the precision of comparative effectiveness estimates. The investigators need to weigh the tradeoffs between bias and precision when they are considering a new-user design.

Objective. Our commentary considers the following question: How important is it for investigators to follow patients from the day they started treatment with the study interventions? Our objective is to start a dialog on the value of the incident user design in comparative effectiveness research.

Methods. We reviewed published case studies in which investigators had used the incident user design or alternative design. We also reviewed methods papers on the incident user design. Our commentary was informed by expert opinion, not systematic evidence.

Contents

Author affiliations:

Eric S. Johnson, Ph.D.[a]
Barbara A. Bartman, M.D., M.P.H.[b]
Becky A. Briesacher, Ph.D.[c]
Neil S. Fleming, Ph.D.[d]
Tobias Gerhard, Ph.D.[e]
Cynthia J. Kornegay, Ph.D.[f]
Parivash Nourjah, Ph.D.[b]
Brian Sauer, Ph.D.[g]
Glen T. Schumock, Pharm.D., M.B.A.[h]
Art Sedrakyan, M.D., Ph.D.[i]
Til Stürmer, M.D., M.P.H.[j]
Suzanne L. West, Ph.D., M.P.H.[k]
Sebastian Schneeweiss, M.D., Sc.D.[l]

[a]Center for Health Research, Kaiser Permanente
[b]Center for Outcomes and Evidence, Agency for Healthcare Research and Quality
[c]Division of Geriatric Medicine, University of Massachusetts Medical School
[d]Center for Health Care Research, Baylor Health Care System
[e]Institute for Health, Health Care Policy, and Aging Research, Rutgers University
[f] Office of Surveillance and Epidemiology, Food and Drug Administration
[g]Salt Lake City IDEAS Center, Veterans Affairs Salt Lake City Health Care System
[h]Department of Pharmacy Practice, University of Illinois at Chicago School of Pharmacy
[i]Division of Outcomes and Effectiveness Research, Weill Cornell Medical College
[j]Department of Epidemiology, University of North Carolina Gillings School of Global Public Health
[k]Health, Social and Economics Research, Research Triangle Institute International
[l]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham & Women’s Hospital, Harvard Medical School

Introduction

Comparative effectiveness research includes cohort studies and registries that lack random allocation of interventions. How important is it to follow patients from the day they initiated treatment with the study interventions? How well can non-randomized studies approximate randomized controlled trials if they follow continuing or prevalent users? What tradeoffs do investigators face when deciding where their study will fall on the continuum from restricting cohort enrollment to patients who are naïve to the entire class of an intervention (most restrictive design) to expanding enrollment to patients regardless of their past use?

Our paper considers the questions outlined above and related issues to start a dialogue on the value of incident user designs in comparative effectiveness research. We take the incident user design as a reasonable default strategy because it protects the evidence from biases, especially when investigators use secondary data sources, such as healthcare databases. Although the incident user design is preferable on theoretical grounds, there may be exceptions where that eligibility criterion does not matter and a less restrictive study design may provide a valid answer that is more timely, more affordable, and more applicable to routine care. We review case studies where investigators have explored the consequences of designing a cohort study by restricting to incident users, but most of the discussion has been informed by expert opinion, not systematic evidence.

The objective of this paper is to consider the incident user design as a default or "first-line" study design for comparative effectiveness research and to provide guidance with discussion on the advantages and limitations of the approach. Although the recommendations could apply to any non-randomized study—regardless of why the data were collected—the recommendations are intended for studies conducted with secondary data sources (i.e., data collected for other reasons).

Issues That Motivate the Incident User Design

Investigators struggle to design non-randomized studies that obtain findings as credible as those from randomized controlled trials. To achieve that credibility, they restrict the question, the design, and the analysis.[1-3] Sometimes the findings from non-randomized studies of comparative effectiveness disagree with the findings from randomized controlled trials. Discrepancies between the cardiovascular findings from the Women's Health Initiative clinical trial and its cohort study of hormone therapy—as well as other studies on the topic—may be the most discussed example.[4-7] Ray suggested that the discrepancy may be explained, in part, by restricting non-randomized studies to new (or incident) users of hormone therapy because the incident user design can reduce biases that occur when comparisons include patients who were already using the drug at the start of the study.[4]

Yet other investigators have approximated the findings from randomized controlled trials by comparing current users of drug therapy and ignoring the duration of drug use. For example, Psaty and colleagues reported that current users of angiotensin-converting enzyme inhibitors and current users of diuretics experienced a similar rate of myocardial infarction: Their case-control study findings (for that comparison and outcome) agreed closely with similar head-to-head comparisons in the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT).[8,9] Ray and colleagues found that the excess rate of coronary heart disease events among patients using COX-2 selective non-steroidal anti-inflammatory drugs was consistent for current users and the subgroup of incident users—a rare comparison of strategies for defining cohorts within a cohort study.[10] Given the inconsistency in these examples, what is the theoretical motivation to prefer the incident user design over other choices that enroll continuing or prevalent users?

Avoiding Adjustment of Intermediate Covariables

One reason to prefer the incident user design is that it avoids the problem of adjusting for characteristics that may be in the causal pathway.[4] For example, in the case-control study noted above, Psaty and colleagues went to exceptional effort to obtain patients' pretreatment blood pressure values, which were documented in paper charts at the health maintenance organization (HMO) an average of 11 years before the index date that determined current drug use.[8] Even with that exceptional effort, pretreatment blood pressure values were missing in one-third of patients who started treatment before joining the HMO. Had Psaty and colleagues adjusted for patients' *most recent* blood pressure values, they would have biased the comparisons between classes of antihypertensive drugs because the most recent blood pressure value is in the causal path between treatment and myocardial infarction. Because most non-randomized studies lack detailed historical data on pretreatment characteristics, it's often more credible to restrict the design to incident users—for whom such pretreatment characteristics can be collected more completely and reliably.

A Fair Representation of Early and Late Events

A second reason to prefer the incident user design is that it captures all events that occurred after the start of therapy.[4,11,12] Molride and Abenhaim explained that some patients are susceptible to harm from a drug and those events may occur earlier in the course of therapy. Once these susceptible patients have suffered events early in the course of therapy, only less susceptible patients remain. Mixing incident and prevalent users may obscure excess harm because the effect measure is weighted toward prevalent users who provide the majority of person-time and were less susceptible to the harm.[12,13]

Guess described a related idea by noting that the hazard ratio for harm changed over time since the start of drug use.[14] Such changes in effect size as a function of the duration of drug use may result from shifts in cohort composition—as described above— or biologic effects or both. Similarly, an intervention's benefits may require an induction period of months or years to reduce clinical event rates: Newly diagnosed patients with diabetes who started intensive therapy for glucose control took up to 10 years to achieve the clinical benefits in the United Kingdom Prospective Diabetes Study.[15] When Prentice and colleagues considered patients' duration of hormone therapy—before the start of the study—in the Women's Health Initiative (WHI) cohort study, their findings approximated more closely the findings from the WHI's controlled trial—at least for some endpoints.[5]

Newly marketed drugs may have a disproportionate share of incident users compared with drugs that were marketed years earlier. For example, a cohort study that was conducted shortly after the marketing of celecoxib that compared patients currently using celecoxib versus those currently using naproxen could distort estimates of comparative effectiveness because the celecoxib users would be more likely to be incident users at higher risk of any early harms related to NSAIDs. Restricting enrollment to incident users is one way to reduce that potential distortion because it enables comparisons at a comparable time in the natural history of their treatment.

Ray traced the idea of the incident user design back to Feinstein's paper on "chronology bias" which appeared in 1971.[16] Ray's own review of "new-user designs" for non-randomized studies remains the most comprehensive account of its value in reducing bias.[4] McMahon and MacDonald provide an earlier and thoughtful consideration of the "new user design".[11] A task force on research practices for retrospective databases organized by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) has also addressed the incident user design and its alternatives.[13] Our paper builds on their efforts by considering how the incident user design applies to comparative effectiveness research. We expand on their previous discussions to consider other interventions, such as medical devices, although most of the examples concern medications. The paper asks that investigators consider the incident user design as a default strategy, but recognizes that there will be exceptions where current user designs may be preferred. The paper encourages more transparent reporting of design choices—in the spirit of STROBE (Strengthening the Reporting of Observational Studies in Epidemiology)—and an appreciation of the tradeoffs that those choices may entail for validity, applicability, timeliness, and feasibility.[17]

Defining Incident Users in Retrospective Studies: Tradeoffs for Internal Validity and Applicability

No consensus exists for defining an incident user of a drug or other intervention using secondary data sources, such as computerized pharmacy fill records or electronic health record prescription orders. Ray and colleagues chose a 365-day window without pharmacy fills for the cohort-defining drug to define incident use of non-steroidal anti-inflammatory drug (NSAID) use.[10] In theory, the physician's order date, or prescription, recorded in some electronic health records, could serve as a more meaningful time-zero for approximating a cohort study version of the intention-to-treat approach used in RCTs.[18] For example, some physicians may dispense product samples along with a prescription; if the patient tolerates the sample then he may fill the prescription, which would appear later as a pharmacy claim. Consequently, the prescription date, possibly documented in an electronic health record, could reflect the start of therapy more accurately.[19]

Recurrent use

Because some health plans will allow patients to fill a 180-day supply of medication, submitted as a single pharmacy claim, briefer windows may misclassify patients who actually used their medication during the baseline period. For example, a patient may have filled his prescription 200 days before the index date selected to define incident use. If the study used a 180-day window to define incident users, he would appear to be an incident user, but would have been taking medication during the months when baseline characteristics were measured. The challenge is more complicated than it may seem: Some patients take long drug holidays and then re-start their medication.[20] Others adhere so poorly—say, every other day—that a prescription intended as a 90-day supply can persist for 180 days. Recognizing the range of possibilities brings a sense of humility about classifying patients as incident users; only a proportion of apparent incident users are truly treatment-naïve as of the index date.

When investigators know that a patient has used the intervention before the window chosen to define incidence use, one option is to stratify comparative estimates so that each group has its own baseline hazard for calculating the hazard ratio that captures comparative effectiveness. For example, patients with a known history of NSAID use more than 365 days before the index date could be compared with other patients who had the same known history. Alternatively, patients with a known history of NSAID use more than 365 days before the index date could be excluded as part of a sensitivity analysis, which may be a reasonable alternative if there are too few patients or events to provide statistically stable, adjusted, stratified comparisons. If the investigators know that medication is used episodically (e.g., cycles of chemotherapy) and they wish to capture the totality of benefits and harms across those episodes, then they should consider more complex structural models that address time-varying drug exposures and confounders; otherwise, it's preferable to evaluate the benefits and harms for the first observed episode only.[21]

Recently Marketed Drugs

For a recently marketed drug, a 365-day window may identify patients for whom the first observed prescription fill in a given database represents their first-ever use (i.e., truly treatment-naïve). For older drugs, a 365-day window may identify patients starting a new *episode* of therapy, but not necessarily their first-ever use. For example, when Ray and colleagues conducted their cohort study using pharmacy data from 1999 through 2001, naproxen has been marketed in the US since 1980; consequently, apparent incident users during 1999 through 2001 may have had past episodes of naproxen use and survived any harms that they experienced during earlier episodes.[2,10] In contrast, apparent incident users of the more recently marketed COX-2 selective NSAIDs, celecoxib and rofecoxib, were more likely to be treatment naïve—at least to those products (but not all NSAIDs). If investigators wish to analyze recurrent episodes of therapy, it's important to adjust standard errors for the correlation of episodes within patients to obtain the correct confidence intervals.[22]

Stricter Definitions of Incident use

Stricter definitions of incident use may improve the internal validity of comparative effectiveness estimates for the reasons outlined in the previous section. But those improvements in validity entail tradeoffs for applicability and precision. For example, patients starting therapy after 365 days without therapy may be at an earlier point in the natural history of their illness (or may be experiencing a milder severity) and therefore at a lower absolute risk of clinical events than patients who would be eligible according to a 180-day or 90-day "wash-out" period, a term borrowed from randomized controlled trial protocols.[23] The extent of any difference in the absolute risk probably depends on the indication and the duration that defined incident use. When Schneeweiss and colleagues calculated the rate of suicide and suicide attempts for all incident users of serotonin reuptake inhibitors (SSRIs), they found that the one year rate was slightly higher in incident users defined by one-year without an antidepressant medication (6.03 per 1,000 person years; 95% CI, 5.54 to 6.55) than in incident users defined by *three*-years without an antidepressant medication (5.18 per 1,000 person years; 95% CI, 4.65 to 5.75).[24]

Requiring patients to have no use of any therapies in the entire *class*—versus requiring patients to have no use of a specific product or intervention—could reduce applicability to a greater extent. Consider the initiation of TNF alpha antagonists in patients with rheumatoid arthritis. Most of these patients used other disease-modifying antirheumatic drugs (DMARDs) and have now switched to a second-line therapy. In this situation, or similar scenarios, stepped-up therapy correlates with progression of the condition. Consequently, investigators should differentiate between comparative effectiveness in first-line therapy versus comparative effectiveness in second-line therapy.[25] For second-line therapy, the cohort would be defined by using a common first-line therapy, say methotrexate in the arthritis example, and the study intervention would be the addition of or switch to a product in the class of TNF alpha antagonists. Such a comparison of incident second-line users would improve the comparability of patients' arthritis severity and progression because they required stepped-up therapy.

Trialists may not need to consider the washout period or window for defining incident use as carefully as investigators conducting non-randomized studies because treatment history is balanced through randomization. For example, ALLHAT did not require any washout period for patients' usual antihypertensive therapies; they switched to their randomly allocated therapy when the trial began.[9] As a counter example, women who wished to participate in the Women's Health Initiative randomized controlled trial were required to undergo a three-month washout period.[26]

Reduced Study Size as a Consequence of Increasing Restrictions

An incident user cohort based on the 365-day definition may produce a less biased estimate of comparative effectiveness, but that finding (e.g., the risk difference) may not apply to as many patients with the condition. Restricting enrollment to incident users can dramatically reduce the size of the cohort and the precision of the comparative effectiveness estimates. For example, when Schneeweiss and colleagues identified elderly patients who filled statin therapies according to a pharmacy claims database, 61,000 met the definition of current use (as of the index date), but only 21,000 patients met the 365-day definition for incident use.[2] Investigators with secondary data sources may find that definitions requiring more than 365 days of "wash-out" exclude another 20% or more of patients: In US insurance plans, 20% of members typically discontinue insurance coverage annually, often switching to a new insurance plan.[27] Stricter definitions of incident use may reduce the precision of comparative effectiveness estimates to a point where the confidence intervals can no longer rule-out clinically important levels of harm or benefit. In some instances, stricter definitions of incident use may require multi-center studies to achieve adequate precision—especially for testing equivalence and non-inferiority hypotheses.[28]

Defining Incident Users in Prospective Studies: Tradeoffs for Internal Validity and Applicability

Prospective Cohort Studies

The Physicians' Health Study was a randomized controlled trial designed to evaluate the benefits and harms of aspirin therapy in relation to preventing cardiovascular events.[29] Investigators continued to follow patients after the trial stopped; Stürmer and colleagues took advantage of the fact that none of those enrolled in the trial were regular aspirin or NSAID users at baseline—an exclusion criterion for the trial—and used the aspirin exposure, along with other NSAID exposure started during follow-up to conduct a prospective cohort study of the relation between incident use of NSAIDs and the incidence of colorectal cancer.[30] As with the retrospective cohort studies discussed above, Stürmer and colleagues chose a 365-day window to define incident use, but based the classification on physicians' self-reported recall of past non-steroidal anti-inflammatory drug use. One of the limitations of prospective cohorts and registries (that lack computerized pharmacy records) is that they often depend on patients' self-reported history of medication use, which may misclassify some patients as incident users when they forgot to report recent medication use. Sometimes that information on medication history is recorded in patients' medical charts, but prospective studies may not have access to those charts unless the study is nested within a health plan. If recent use of a medication is an exclusion criterion for a prospective study, then it could be expensive to screen potentially eligible patients to identify a sufficient number of incident users. For example, the Women's Health Initiative prospective cohort study found that 33% of the 53,000 eligible women were current users of combined estrogen-plus-progestin preparations at baseline.[5] Had the WHI investigators tried to assemble a cohort of incident users, thousands of interviewed women would have been excluded at great expense to the study. That expense is trivial in retrospective studies with computerized pharmacy records (or electronic prescription records) because patients do not need to be recruited and screened through interviews; a computer algorithm can efficiently query millions of patient records.

Specifics About Registry Studies

A prospective product registry is often a cohort study and can include among its objectives evaluating the comparative effectiveness of interventions.[31] All of the issues outlined above for prospective cohort studies apply to registries, too. Although some registries enroll patients starting therapy (i.e., incident users) and follow them for outcomes from that date forward, other registries enroll patients who started therapy previously and try to capture early outcomes, such as harms, through chart reviews. For example, the British Society for Rheumatology Biologics Register enrolled patients who had started either etanercept or infliximab therapy within the previous six months (i.e., before enrollment in the registry).[32] It's unclear whether the earliest months of anti-TNF therapy contributed to the analysis (i.e., incident users) or whether the follow-up began *after* the start of therapy (i.e., prevalent users). If the earliest months of anti-TNF therapy contributed to the analysis, the study could suffer an "immortal time bias."[33] The bias occurs because some patients had to tolerate the anti-TNF therapy—and survive any early harms—in order to contribute follow-up to the analysis: The complete clinical story may be missing from the registry's findings and comparative findings may be distorted. Because some registries are product-specific, they may be limited in their suitability for comparative effectiveness research; it may not be possible for investigators to find a registry of incident users of alternative interventions. Historical registries, sometimes started before the marketing of a new drug, may offer one option despite some limitations.

Specifics About Medical Device Registry Studies

Medical device registries also need to consider the design choices related to incident users with subtle distinctions that merit elaboration. For example, should the date of incident use be defined in relation to the medical procedure or the date of the decision to opt for the device? It's analogous to the tradeoffs outlined above for the prescription fill date vs. the prescription date. In theory the earlier date that marks the decision to opt for the device may be preferable because clinical characteristics known to the physician as of that date may capture preferential treatment more accurately and put investigators in a better position to control possible confounding by indication. In practice, that date may not be known unless investigators have access to the physicians' notes recorded in the chart. Patients who do not have the procedure—despite an earlier, documented decision to have the procedure—contribute to misclassification and can make it harder for investigators to identify benefits and harms. If the comparison is between a device and a medical therapy, then it would be important to select a comparable time-zero date for the start of follow-up. For example, investigators may consider choosing the decision date for the device and the prescription date for the drug to ensure comparable measurement of baseline characteristics that may predict treatment and to start counting events from the same point in patients' natural history. The distinction matters when the procedure is delayed but the prescription is filled quickly. Because most retrospective databases don't have the prescription date, it would be preferable to chose the prescription fill date and the procedure date and accept the non-differential bias.

Recommendations for Reporting for Sections 2 and 3

(1) Investigators should report whether they designed the study to compare incident users of the intervention or whether they enrolled patients according to different eligibility criteria. Those eligibility criteria and their effect on the numbers of patients should be documented in a CONSORT-style participant flow diagram that will allow readers to assess the applicability of the findings to their populations and settings.[34]

(2) Investigators should report how they defined the dates of intervention use, and for studies that followed incident users, the window of time used to classify patients as incident users along with the clinical rationale for that window of time.

(3) Investigators should report whether the baseline characteristics (covariables) were measured before incident drug use or whether those characteristics may reflect the effects of the study-defining interventions. For characteristics measured before incident drug use, investigators should report the timing of those measurements in relation to the start of the intervention.

(4) Investigators should conduct sensitivity analyses with varying durations of the washout period to illustrate the stability of findings with respect to validity and precision.

Advantages of the Incident User Design for Improving Internal Validity

Improved Confounder Control

One of the advantages of the incident user design is better control of confounding because patients using the intervention or its comparator are both initiating a new course of treatment in routine care. That means both patient cohorts were seen by physicians who evaluated their condition and decided that their condition warranted treatment. Patients then took the initiative to fill the prescription or undergo the procedure. All of these characteristics--a mixture of diagnostic skill, treatment guidelines, and medical sociology--indicated that the patients in the intervention and comparator cohorts were more similar with respect to their disease state than the entire population with the condition.

Avoidance of Intermediate Variables

The incident user design also dramatically reduces the opportunity for investigators to adjust for variables that are in the causal pathway by ensuring that all patient characteristics were measured before treatment initiation and before follow-up started. For example, if a cohort study compared the effectiveness of antihypertensive therapies, the investigators might want to adjust for baseline blood pressure values if newer drugs were preferentially prescribed to patients with worse hypertension. If the cohort's baseline blood pressure values were measured while patients were taking their cohort-defining drugs, the cohorts might appear more comparable at time zero than they were at the time the drugs were prescribed. If one drug is more effective than another at preventing events, that effectiveness might be mediated through superior blood pressure control; controlling for treated values would obscure the true difference in effectiveness. Similarly, if a cohort study enrolled current users and compared the rate of cardiovascular events for COX-2 therapies versus naproxen, adjusting for treated "baseline" blood pressure values could bias the rate ratio estimate toward 1.0: An increase in hypertension may be part of the mechanism through which some COX-2 therapies increase the rate of cardiovascular events.

Adequate Consideration of Time Since Marketing and Time Since Initiation

Incident user designs add value to the comparative effectiveness evidence for another reason: accurate reflections of the induction period or time-to-event. When time-zero and follow-up are aligned with the incident use of an intervention and its alternative, investigators can compare cohorts' times until event and obtain a valid hazard ratio. That's not necessarily true when investigators conduct the same analysis with prevalent users because the induction period may have started during the baseline period or earlier. For example, unless all of the drug cohorts have a similar distribution of start times before time zero, the comparisons may be confounded. If newer drugs were started more recently, the newer drugs may appear safer than older drugs because their apparent rate would be lower; investigators would observe a higher event rate for older drugs because much of their induction period would be obscured by the baseline period.

By following patients from the time they start therapy—the incident user design—investigators can identify all of the outcomes that may be related to the therapy, including therapy discontinuation. Prevalent user designs may miss some early events, especially treatment discontinuations, because of problems with tolerability or poor response. Following patients from the time they started therapy protects against immortal time bias: Patients need not survive through the early months of therapy and persist with therapy to be eligible for the study, which can distort the absolute event rates and the estimates of comparative effectiveness.[33] Another advantage of the incident user design is that it can capture the clinical consequences of an entire therapeutic strategy including co-interventions and dose titration that may be important for understanding effectiveness under routine practice conditions (i.e., pragmatic questions), instead of focusing on the narrower efficacy or explanatory question for the subset of patients who adhere with therapy, are stable on their dose, etc.

The Incident User Design Complements Propensity Score Analyses

The incident user design results in more effective propensity scores because the baseline characteristics that contribute to the score predict incident events (e.g., a fill at the pharmacy). When the propensity score predicts prevalent use as, it's harder to interpret its meaning and the propensity score may not be as effective at reducing confounding. For example, some propensity scores predict a combination of incident drug use, persistence with a drug started months (or years) earlier, and possibly drug switches within a class. Different characteristics may predict each of those endpoints more accurately than trying to model them as a composite endpoint of current use on the index date. Another advantage of developing propensity scores to predict incident drug use is that it reminds investigators to model a choice among therapies instead of treating the lack of therapy as if it were meaningful inception date. For example, Schneeweiss and colleagues developed a propensity score to predict incident use of conventional or atypical antipsychotic therapies.[35] It's harder to assign a meaningful incidence date for an intervention that did not happen; however, investigators may be able to assign a date when patients started an entirely unrelated class of medications—instead of an alternative treatment for the same indication.[2]

Tradeoffs of the Incident User Design

Reduced Study Size and Reduced Precision

Although the incident user design offers the many advantages explained above, the most obvious tradeoff is a loss of precision in estimates of comparative effectiveness. The numbers of incident users may be too low in some databases, for example, not-for-profit HMOs that tend to adopt innovative drugs more cautiously than publicly-traded insurance plans. In that scenario, a multi-site study may be required to obtain a sufficient number of incident users. In some instances, the numbers of incident users available even after pooling available databases may be too limited to justify the study. That scenario would constrain the timeliness of the evidence for decision-makers.

A related point is that data sources may include too few events—benefits or harms—even if they include a sufficient denominator of patients. Two concerns confront the investigator when there are few events. First, the confidence intervals for the estimates of comparative effectiveness may be so wide that they fail to exclude clinically important benefits or harms (say, a 50% excess rate); the findings would remain inconclusive in relation to the study's hypothesis. Second, there may be enough events for precise confidence intervals, but too few events for the number of covariables that require statistical adjustment. Unless the investigators are willing to use shrinkage estimation methods (or other sophisticated alternatives), the hazard ratios or other effect measures may suffer a "sparse data bias" that can inflate the estimate of benefit or harm; propensity scores, risk scores and other variable-reduction methods can help address that problem.[36]

Reduced Ability to Study Long-Term Effects

Another tradeoff in the incident user design is that it may not allow evaluation of the long-term effects of drugs, such as cumulative years of exposure. Among incident users, it's not uncommon for half of the patients to discontinue by the first year of follow-up. Relatively few patients will have five or more years of continuous exposure to a drug, especially a newly marketed drug. Moreover, many databases are limited in their ability to follow patients for five or more years because patients (or their employers) discontinue insurance coverage with one health plan and the continuing drug use cannot be linked to their new health plan. This limitation of the incident user design probably impacts benefit events more than harm events because the induction period for meaningful risk reduction may only be seen after years of adherence to therapy.

Reduced Generalizability

The other main tradeoff that results from choosing to study incident users is a reduction in applicability or generalizability. The estimates of comparative effectiveness may be more valid (internally), but apply to fewer patients. In many cases, the patients for whom decision-makers require comparative evidence are those patients already using one of the drugs in a class; requiring all patients in the cohort to have a 365-day window without use of any drugs in the class may focus attention on patients at an earlier stage in their natural history or with a lower disease severity. The longer the window used to define incident use, the more likely that patients' absolute risk of the events for benefits or harms will be low—and the evidence may not apply to more typical patients already taking such therapies in the community setting.

One related challenge is to evaluate the comparative effectiveness of drugs that should be used as second-line therapy. The valid comparison requires another drug or drugs that should be used as second-line therapy: Incident use in this context would only refer to the second drug (possibly an augmentation); it would be acceptable for patients to have continuing use of the first-line drug during the baseline period. In the community setting, the drugs may not be used in that preferred sequence (e.g., according to clinical practice guidelines); consequently, the most valid estimates may apply to relatively few patients of interest to decision-makers.

Finding the Right Comparison Cohort

The identification and justification of the most appropriate comparison group is a challenging but generic issue to all clinical studies. Using comparator interventions with the same indication can reduce confounding substantially.[37] Similarly, investigators' ability to identify an effect of an intervention depends on fluctuations in the natural history of patients' condition, which generates statistical noise.[38] Because interventions are more likely to be started at comparable times during patients' natural history—for example, during a worsening of symptoms—the incident user design may reduce confounding; such an advantage is lost when incident users of one intervention are compared with non-users or prevalent users of "usual care" with no defined onset. Although some pragmatic, randomized controlled trials compare an intervention to usual care, it may be more complicated to identify the onset of usual care cohort in retrospective databases. Because usual care practice patterns may exhibit greater heterogeneity in routine practice than in a pragmatic trial with a protocol, investigators must take extra care in defining usual care in retrospective databases so that decision-makers can interpret the meaning of the comparisons.

For drugs that are available as combination products, it is important to handle the baseline period uniformly across comparisons: If patients are switching from monotherapy to a fixed-dose combination product, then the comparator would need a similar sequence. But if patients are starting a fixed-dose combination product *without* having tried monotherapy with one of the drugs in the combination, then comparisons should seek patients who are similarly new to combination therapy (e.g., filling two different classes of drugs on the same day). For example, if investigators wanted to evaluate the comparative effectiveness of a fixed-dose asthma therapy that includes a long-acting beta-agonist and an inhaled corticosteroid, they would need to consider how patients arrived at that regimen and restrict the comparisons accordingly. Otherwise, patients who started on an inhaled corticosteroid according to guidelines and then stepped-up to a long-acting beta-agonist would have baseline characteristics that were modified by treatment and it would be difficult to evaluate possible confounding.

It is one of the basic assumptions of comparative effectiveness research that two active agents or treatment strategies will be compared with each other. If a drug is the first in its class, it can still be compared to the standard of care prescribed before this new drug became available; however, that comparison may introduce confounding. The incident user design has inherent limitations when comparing an intervention to non-users. Non-users are different from patients using placebo as they may not have the indication for use or may have contraindications. Moreover, non-use of an intervention may be a marker for inadequate access to the health care system: Non-users may be less likely to visit a physician and pay to fill a prescription. In non-randomized studies, non-users are often fundamentally different patients in ways that are difficult to measure in secondary data--and sometimes in primary data. In retrospective data sources, non-user comparisons may generate immortal time bias because investigators sometimes require that patients not use the study intervention during follow-up to avoid treatment cross-over and bias toward the null.[33] Because non-users must survive this period of time, by definition, bias arises. The bias can be easily remedied by starting follow as soon as the non-use status was determined and then censoring patients when they switched from non-use to use of an intervention. Many experienced investigators have inadvertently introduced immortal time bias by comparing use of an intervention to non-users.

Incident User Designs in the Context of Other Design and Analysis Choices

Part of the value of the incident user design results from the other design and analysis options that it allows, which may improve the consistency between findings from non-randomized studies and randomized controlled trials.[2] For example, the incident user design helps investigators confront the question of the most rigorous control cohort: Should incident users of one drug or class be compared with incident users of another drug or class? Or can valid estimates of comparative effectiveness be obtained by comparison with non-users? If so, what's the relevant time-zero for non-users? Ray and colleagues compared incident users of statin medications to two control cohorts to evaluate their benefits in reducing the risk of hip fracture.[39] When they compared incident statin users to non-users, statins appeared to reduce the risk of fracture by approximately 40%. When they compared incident statin users to incident users of other lipid-lowering drugs (i.e., active controls), the risk of fracture *increased* by approximately 40%. We lack evidence to know how much active control cohorts matter for estimates of comparative effectiveness. But the incident user design at least prompts a conversation about the most valid control cohort—even if it is a sensitivity analysis.

Incident user designs allow investigators to undertake more meaningful propensity score analyses that predict the start of therapy—instead of a composite endpoint that may include persistence or switching for some patients. The choice to incorporate propensity scores into an analysis of comparative effectiveness is separate from the choice to undertake an incident user design, but the two strategies work well together. By comparing the degree of overlap in propensity scores for two interventions, investigators can spot problems with clinical equipoise that may bias the estimates: Some patients may have an unacceptably low probability of treatment that should disqualify them from the comparison; other patients may have a near-certain probability of treatment that should disqualify them.[40] The incident user design helps investigators frame questions of treatment choice and clinical equipoise. Studies that opt for (event) risk scores instead of propensity scores to control confounding would also benefit from the incident user design.

Incident user designs can add value to comparative effectiveness studies by allowing investigators to define the drug cohort according to use at time-zero and carrying that exposure status forward in the analysis (without updating that exposure status when patients discontinue therapy or adhere poorly). Although cohort studies and registries lack randomization, an analysis based on patients' prescription fill at time-zero better approximates the intention-to-treat analysis advocated by trialists, which ignores patients' actual use of the intervention. For example, Schneeweiss and colleagues compared mortality rates over a 180-day period in relation to incident use of conventional or atypical antipsychotic therapies filled at time zero—regardless of whether patients refilled those drugs, appeared to switch classes of medication, etc.[35] The problem with updating patients' drug cohort in relation to their pattern of adherence is that it may confound estimates of comparative effectiveness and investigators lack information on the time-varying characteristics that would help them understand why some patients adhered at any given time and others did not.[41] For example, Brookhart and colleagues found that patients who refilled their statin prescription were more likely to seek preventive services than patients who only filled their statin prescription once.[41] Although investigators could use patients' current drug use at time zero to define cohorts, much of the healthy-user or adherence bias would be present and could distort estimates of comparative effectiveness. Questions remain on the preferred date to serve as time-zero: Should time-zero start with the prescription or the prescription fill at the pharmacy (or analogous dates for decisions about devices and their procedures)? Investigators using secondary data sources may not have a choice (e.g., claims data may not reflect the date of the prescription). But investigators using prospective sources should weigh the tradeoffs of increasing misclassification by using the earlier date (because some patients will not get the intervention) versus increasing confounding by using the later date (because investigators may not know why some patients followed-through with the intervention).

Priority Research Gaps for Understanding Tradeoffs in Defaulting to the Incident User Design

(1) Quantify the effects of varying durations of washout periods on validity and precision in several example studies and databases and investigate the external factors that influence the findings (e.g., median time between refills).

(2) Evaluate first-line, second-line, and third-line therapy defined as incident users and compare their findings with continuing or prevalent users

(3) Evaluate potential tradeoffs between including (following) all drug initiators vs. including only regular users (e.g., after 2-script run-in period).

(4) Understand how the tradeoffs of the incident user design apply to interventions other than medications because most of the case studies have compared medications with other medications (instead of surgical procedures, etc).

References

1. Vandenbroucke JP. When are observational studies as credible as randomized trials? Lancet 2004;363:1728-31.

2. Schneeweiss S, Patrick AR, Stürmer T, et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. Medical Care 2007;45(Suppl 2):S131-42).

3. Perrio M, Waller PC, Shakir SAW. An analysis of the exclusion criteria used in observational pharmacoepidemiological studies. Pharmacoepidemiol Drug Safety 2007;16:329-36.

4. Ray WA. Evaluating medication effects outside of clinical trials: new user designs. Am J Epidemiol 2003;158:915-20.

5. Prentice RL, Langer R, Stefanick ML, 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;162:404-14.

6. Petitti DB, Freedman DA. Invited commentary: How far can epidemiologists get with statistical adjustment? Am J Epidemiol 2005;162:415-18.

7. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiol 2008;19:766-79.

8. Psaty BM, Heckbert SR, Koepsell TD, et al. The risk of myocardial infarction associated with antihypertensive drug therapies. JAMA 1995;274:620-25.

9. ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs. diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). JAMA 2002;288:2981-997.

10. Ray WA, Stein CM, Daughtery JR, et al. COX-2 selective non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease. Lancet 2002;360:1071-73.

11. McMahon AD, MacDonald TM. Design issues for drug epidemiology. Br J Clin Pharmacol 2000;50:419-25.

12. Moride Y, Abenhaim L. Evidence of the depletion of susceptibles effect in non-experimental pharmacoepidemiologic research. J Clin Epidemiol 1994;47:731-737.

13. Cox E, Martin BC, Van Staa T, et al. Good research practices for comparative effectiveness research: Approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: The International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force—Part II. Value in Health 2009;12:1053-61.

14. Guess HA. Exposure-time-varying hazard function ratios in case-control studies of drug effects. Pharmacoepidemiol Drug Safety 2006;15:81-92.

15. UKPDS Group. Intensive blood-glucose control with sulfonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet 1998;352837-53.

16. Feinstein AR. Clinical biostatistics. XI. Sources of “chronology bias” in cohort statistics. Clin Pharmacol Ther 1971;12:864-79.

17. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. PLoS Med 4:e297.doi:10.1371/ journal.pmed.0040297.

18. Jacobus S, Schneeweiss S, Chan KA. Exposure misclassification as a result of free sample drug utilization in automated claims databases and its effect on pharmacoepidemiologic studies of selective COX-2 inhibitors. Pharmacoepidemiol Drug Safety 2004;13:695-702.

19. Hippisley-Cox J, Coupland C. Unintended effects of statins in men and women in England and Wales: population-based cohort study using the QResearch database. BMJ 2010;340:c2197 doi:10.1136/bmj.c2197.

20. Vrijens B, Vincze G, Kristanto P, et al. Adherence to prescribed antihypertensive drug treatments: longitudinal study of electronically compiled dosing histories. BMJ 2008;336:1114-7.

21. Hernán MA, Cole SR, Margolick J, et al. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. Pharmacoepidemiol Drug Saf 2005;14:477-91.

22. Stürmer T, Glynn RJ, Kliebsch U, et al. Analytic strategies for recurrent events in epidemiologic studies: background and application to hospitalization risk in the elderly. J Clin Epidemiol 2000;53:57-64.

23. Knipschild P, Leffers P, Feinstein AR. The qualification period. J Clin Epidemiol 1991;44:461-4.

24. Schneeweiss S, Patrick AR, Solomon DH, et al. Variation in the risk of suicide attempts and completed suicides by antidepressant agents in adults: a propensity score-adjusted analysis of 9 years' data. Arch Gen Psych 2010;67:497-506.

25. Solomon DH, Lunt M, Schneeweiss S. The risk of infection associated with TNF antagonists: making sense of epidemiologic evidence. Arthritis Rheumatism 2008;58:919-28.

26. Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy post-menopausal women. Principal results from the Women's Health Initiative randomized controlled trial. JAMA 2002;288:321-33.

27. Short PF, Graefe DR, Schoen C. Churn, churn, churn: how instability of health insurance shapes America's uninsured problem. Issue brief, The Commonwealth Fund. New York, NY 2003.

28. Fleming TR. Current issues in non-inferiority trials. Statist Med 2008;27:317-32.

29. Steering Committee of the Physicians' Health Study Research Group. Findings from the aspirin component of the ongoing Physicians' Healthy Study. N Engl J Med 1989;321:129-135.

30. Stürmer T, Buring JE, Lee IM, et al. Colorectal cancer after start of nonsteroidal anti-inflammatory drug use. Am J Med 2006;119:494-502.

31. Gliklich RE, Dreyer NA, eds. Registries for Evaluating Patient Outcomes: A User's Guide. AHRQ Publication No. 07-EHC001-1. Rockville: Agency for Healthcare Research and Quality; 2007.

32. Hyrich KL, Watson KD, Silman AJ, et al. Predictors of response to anti-TNF-α therapy among patients with rheumatoid arthritis: results from the British Society for Rheumatology biologics register. Rheumatology 2006;45:1558-65.

33. Lévesque LE, Hanley JA, Kazouh A, et al. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ 2010;340:907-911.

34. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332 doi:10.1136/bmj.c332.

35. Schneeweiss S, Setoguchi S, Brookhart A, et al. Risk of death associated with the use of conventional versus atypical antipsychotic drugs among elderly patients. CMAJ 2007;176:627-32.

36. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. Am J Epidemiol 2008;167:523-9.

37. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005;58:323-37.

38. Glasziou P, Chalmers I, Rawlins M, et al. When are randomised trials unnecessary? Picking signal from noise. BMJ 2007;334:349-51.

39. Ray WA, Daugherty JR, Griffin MR. Lipid-lowering agents and the risk of hip fracture in a Medicaid population. Injury Prev 2002;8:276-9.

40. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. 2006;98:253-9.

41. Brookhart MA, Patrick AR, Dormuth C, et al. Adherence to lipid-lowering therapy and the use of preventive health services: An investigation of the healthy user effect. Am J Epidemiol 2007;166:348-54.