



# Registry Data in Injury Research: Study Designs and Interpretation

Stephen J. Mooney<sup>1,2</sup> · Andrew G. Rundle<sup>3,4</sup> · Christopher N. Morrison<sup>3,4,5</sup>

Accepted: 5 October 2022 / Published online: 20 October 2022  
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2022

## Abstract

**Purpose of Review** Injury data is frequently captured in registries that form a census of 100% of known cases that meet specified inclusion criteria. These data are routinely used in injury research with a variety of study designs. We reviewed study designs commonly used with data extracted from injury registries and evaluated the advantages and disadvantages of each design type.

**Recent Findings** Registry data are suited to 5 major design types: (1) Description, (2) Ecologic (with ecologic cohort as a particularly informative sub-type), (3) Case–control (with location-based and culpability studies as salient subtypes), (4) Case-only (including case-case and case-crossover subtypes), and (5) Outcomes.

**Summary** Registries are an important resource for injury research. Investigators considering use of a registry should be aware of the advantages and disadvantages of available study designs.

**Keywords** Injury · Registry · Research methods · Secondary data · Study design

## Introduction

The circumstances surrounding injury events display underlying patterns that make it possible to identify common causes and outcomes that can be intervened on to prevent future injuries or minimize injuries' impact. For example, the finding from the 1950s that alcohol consumption is a frequent contributor to motor vehicle fatalities provided evidence to support laws outlawing driving while intoxicated [1]. More broadly, injury prevention and control research has made substantial contributions to reduce motor vehicle

fatalities; to prevent falls, burns, and recreational injuries; and to understand injury's persistent links to substance use, and to violence and self-harm [2].

However, injury events are rare, and it can be challenging for researchers to identify antecedent contributing factors prospectively. Injuries are thus well suited to registry-based research, and indeed, substantial resources are committed to recording injuries in registries, including the National Trauma Data Bank (NTDB), Fatality Analysis Reporting System (FARS), and National Violent Death Reporting System (NVDRS). Table 1 briefly describes three registries commonly used in US-based injury research.

Formally, a register is the file of data containing all cases of a health-related condition, and a registry is the corresponding system of registration [3]. Registries are therefore a census of 100% of known cases that meet the inclusion criteria. Inclusion is commonly temporally and geographically bounded (e.g., all eligible cases within the USA after January 1, 2000), although in some instances a spatial criterion is replaced by alternative markers (e.g., attendance at specific hospitals). The inclusion criteria typically drive the definition of the units, which are represented as rows in a dataset. For example, a registry of emergency department admissions will have each admission as a unit and individuals can be represented multiple times if they have multiple eligible hospital admissions.

This article is part of the Topical Collection on *Injury Epidemiology*

✉ Stephen J. Mooney  
sjm2186@uw.edu

<sup>1</sup> Department of Epidemiology, University of Washington, Box 357236, Seattle, WA 98195, USA

<sup>2</sup> Harborview Injury Prevention and Research Center, University of Washington, Seattle, WA, USA

<sup>3</sup> Department of Epidemiology, Columbia University Mailman School of Public Health, New York, NY, USA

<sup>4</sup> Center for Injury Science and Prevention, Columbia University, New York, NY, USA

<sup>5</sup> Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, VIC, Australia

**Table 1** A selection of US-based registries used in injury research

Registry name	Maintainer	Years of data	Inclusion criteria
National Emergency Medical Services Information System	National Highway Transport Safety Administration's Office of EMS & the University of Utah	2009–	Records of emergency medical services responses from US states and territories
Fatality Analysis Recording System	National Highway Transport Safety Administration	1975–	Records of all collisions involving a motor vehicle in which at least one person died as a result of the collision within 30 days
National Violent Death Reporting System	Centers for Disease Control and Prevention	2003–	Records of all violent deaths in the USA

EMS, Emergency Medical Response

However, the natural unit is not always clear. For example, registries of motor vehicle crashes can be separated into victim-, vehicle-, and crash-level units [4], with salient information available at each level. Notably, administrative records that are not actively registering specific conditions (e.g., electronic health record databases) are not included in this definition. For the purposes of this document, we will follow colloquial usage, in which “registry” is used to refer to the data file itself.

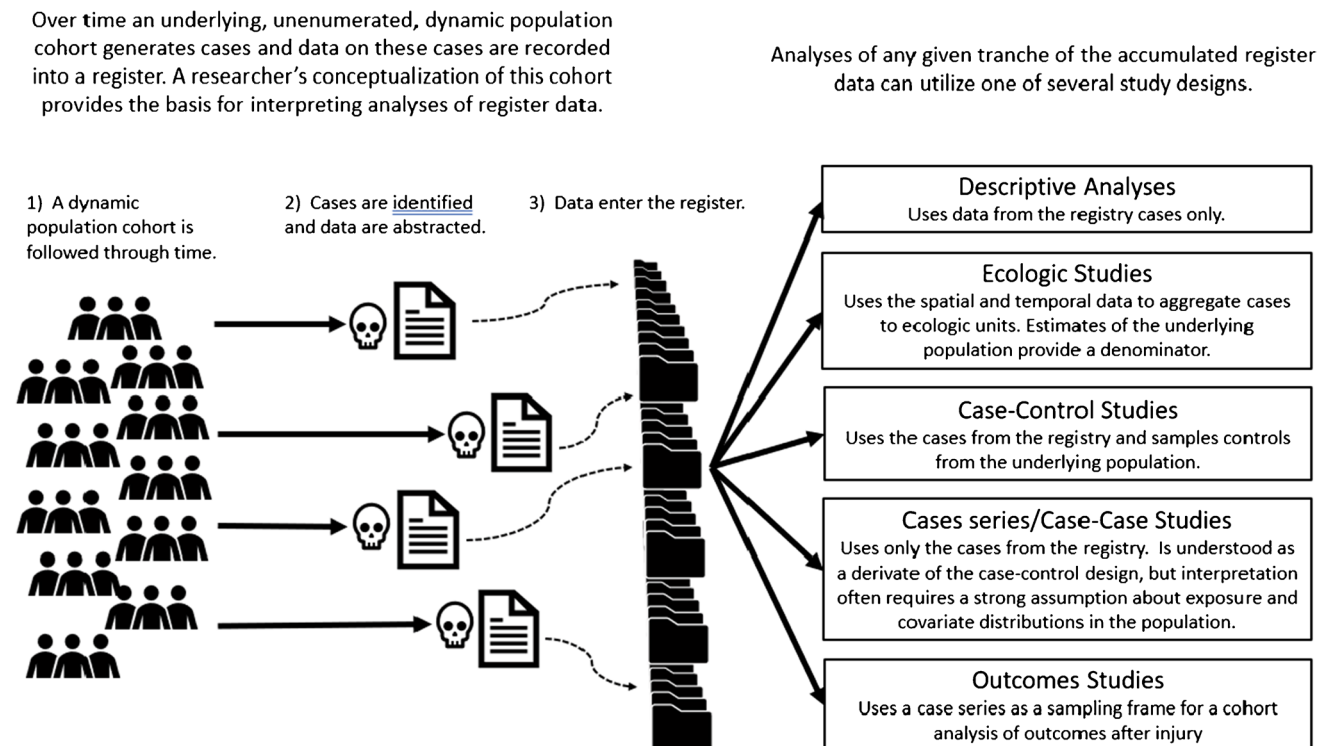
Once data are collected into registries, these registries can be used to study injury using many designs (Fig. 1). Five classes of study designs and several subtypes of those designs that often draw from registries are presented in Table 2.

Over time an underlying, unenumerated, dynamic population cohort generates cases and data on these cases are recorded into a register. A researcher's conceptualization of this cohort provides the basis for interpreting analyses of register data.

In this review, we will briefly discuss each of these distinct uses for registry data for injury research, including pros and cons of each approach, providing examples of their use.

## Descriptive Epidemiology

The most analytically straightforward use of injury registries is to describe the spatial, temporal, and subpopulation distribution of injuries that meet some specified inclusion criteria. These measures of injury frequency can describe distributions within subpopulations sharing a characteristic such as sex or age [12], depict geospatial distributions (e.g., identifying clusters or hot spots) [13], or explore changes



**Fig. 1** Schematic of registry data being used to study injury

**Table 2** Study designs commonly used with injury registry data

Study design	Important sub-types	Key comparison	Key weakness	Example drawing from FARS data
Descriptive epidemiology Ecologic	Cross-sectional ecologic	None	No causal interpretation of results	Traffic safety facts; pedestrians [5]
	Time series ecologic	Differences between aggregated units	Ecologic fallacy; modifiable areal unit problem	Helmet laws and motorcycle rider death rates [6]
Case-control design	Ecologic cohort	Differences within aggregated units	Ecologic fallacy; modifiable temporal unit problem	Maximum blood alcohol concentration laws and fatal alcohol-involved crashes [7]
	Ecologic case-crossover	Differences within and between aggregated units	No individual-level interpretation	“Complete Streets” and adult bicyclist fatalities [8]
	Person-based case-control	Differences within aggregated units, with inclusion determined by case status	Space-time units must be dichotomized with respect to outcomes	
	Location-based case-control	Cases with population-based controls drawn from another dataset	Difficult to identify controls	Drug use and fatal motor vehicle crashes: a case-control study [9]
	Culpability	Places where events occurred and places where they could have but didn't	Inability to assess person characteristics, challenges identifying control locations	
Case series/case-only designs	Case-case	Drivers responsible for collisions and drivers not responsible for collisions	Challenges assigning responsibility for a collision	Role of alcohol and marijuana use in the initiation of fatal two-vehicle crashes [10]
	Case-crossover	One subset of cases compared to another subset of cases	Few scientifically valid scenarios	Determinants of seat belt use among fatally injured teen drivers [11]
Outcomes cohort		Case exposure at an index injury to exposure for the same case at another time	Requires exposure data at time other than injury	
		Outcomes of individuals affected by an injury	Inability to assess prevention	

Abbreviations: *FARS*, Fatality Analysis Reporting System; *NHTSA*, National Highway Transport Safety Administration

over time [14]. Descriptive studies may be used to identify subpopulations bearing a disproportionate burden of disease, but do not attempt to quantitatively assess associations between possible causes and injury incidence.

The primary advantages of descriptive studies are their overall simplicity and the low likelihood that analytic artifacts can be responsible for findings. The primary disadvantage of these studies is that, without a causal focus to the analysis, findings typically cannot guide specific individual or policy changes to prevent the injuries—that is, because descriptive studies do not tell us about impacts of potential causes of injuries, their results should be used for resource allocation and hypothesis generation, but not to identify or select interventions at the individual or population level.

For example, a descriptive finding that pedestrian fatalities have increased over time could be used to build ideas about the possible reasons behind the observed rise, or to allocate resources to interventions already known to reduce fatalities. The finding should not, however, be used alone to advocate directly for specific interventions to reverse the trend, such as further enforcement of distracted driving laws, because whether distracted driving contributed to the rising trend and whether enforcement would reverse the trend cannot be determined from a descriptive analysis.

Key issues for investigators considering this design include appropriate and effective communication of results—because descriptive studies are typically more accessible to non-specialists than more causally focused study designs, technical language, visualizations understood only by experts in the field, and lack of attention to caveats may result in incorrect interpretation by the broad audience of such studies. Additionally, investigators must understand

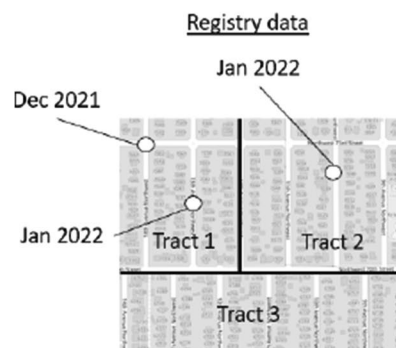
the process that leads to a case being registered. Registries typically aim to comprise a census of cases in their catchment area, but artifacts in referral or ascertainment may impose selection pressures on investigators that could lead to biased estimates of outcome distributions.

For example, the National Highway Transportation Safety Administration’s yearly report summarizing the descriptive epidemiology of pedestrian fatalities from recent FARS data [5] allows researchers and policymakers to track trends in pedestrian safety and helps researchers generate hypotheses about factors that might be affecting pedestrian fatality rates. Similarly, Hemenway used NVDRS data to describe the distribution of homicide committed by children to generate hypotheses regarding possible causes of homicide within this key sub-population [15].

## Ecologic Designs

A second use for registry data starts like a descriptive study—aggregating individual-level registry entries into larger groups—but goes on to assess potential causes of the variation in injury rates between these groups. Aggregation is typically performed within space–time units, such that injury events are combined as counts or rates within units that are bounded by space (e.g., cities, states) and by time (e.g., days, months). These aggregated data can be collapsed into analytic datasets that capture variation by space alone (i.e., an ecological cross-sectional dataset) or time alone (i.e., an ecological time series dataset); or datasets that capture variation by space *and* time (i.e., an ecological panel dataset) (Fig. 2).

**Fig. 2** Visualization of aggregation units and designs for address-level road traffic crash data



### Ecological cross-sectional dataset

Tract	Collisions
1	2
2	1
3	0

### Ecological longitudinal dataset

Date	Collisions
Dec 2021	1
Jan 2022	2

### Ecological panel dataset

Date	Tract	Collisions
Dec 2021	1	1
Dec 2021	2	0
Dec 2021	3	0
Jan 2022	1	1
Jan 2022	2	1
Jan 2022	3	0

The researcher can combine these spatially and/or temporally varying aggregate measures of injury incidence with ecological measures of social, physical, economic, or policy environments to assess associations between these exposures and the injury outcome. The appropriate study design and statistical analytic method will depend on the structure of the available data, the distribution of the outcome measures, and the nature of the exposure. For example, ecological cross-sectional data can be used to compare injury incidence between locations, and ecological time-series data can be used to assess possible determinants of change over time. Two-dimensional panels can accommodate binary exposures in a difference-in-difference framework that allows simultaneous assessment of treatment selection and global time trends to isolate treatment effects. Advances in statistical methods allow researchers to rigorously control for spatial and temporal dependencies, and for both time-fixed and time-varying confounding by place [16–18](p).

Whereas the ecologic design has been widely (and rightly) criticized for misapplication and misinterpretation, it can be very useful in select scenarios. In particular, some causal phenomena operate at the ecological level, so the appropriate unit of analysis is ecological units [19]. For example, the availability of rideshare services (e.g., Uber, Lyft) can change mobility at a population level, including contributing to lower overall motor vehicle ownership in some US cities [20]. Studies of ridesharing access and road traffic crashes should therefore be conducted within ecological units, rather than among individuals who happen to be travelling at a given time. Another instance when ecological designs are advantageous is when this third, panel design (sometimes called the ecological cohort) is used. With panel data, approaches such as a difference-in-differences [21] or synthetic control design [22] can be used to isolate the impacts of specific policies from space-specific or time-specific confounding, providing stronger causal evidence than cross-sectional ecologic studies. Statistical and methodological efficiencies can be achieved using an ecologic case-crossover design, which compares ecological units where an outcome occurs to the same unit at a different time, though this approach requires units to be dichotomized with respect to the outcome, which may not be possible for large space–time units where outcome events are common.

Advantages of the ecologic approach include their ease of development and clear link to group-level policy. Disadvantages of the design include that aggregation loses information about individual experiences that an individual-based study could retain, and that findings are only interpretable at the specific group level studied (e.g., counties) and can be misleading when applied to individuals or other group levels (e.g., states) [23–25].

There are several key issues for the ecologic design. First, investigators must identify a spatial and temporal

scale consistent with their causal theory—for example, green space remediations are hypothesized to affect crime and violence in close proximity to the remediated lots [21], so an analysis at too large a geographic scale (e.g., ZIP codes, municipalities) might fail to identify true effects. A second concern is that registries are frequently deidentified for public use, so the location data about cases, necessary to assign the cases to ecologic units, may be suppressed to prevent subject identification. Finally, even if the spatial and temporal scales are defined appropriately, results depend on the spatio-temporal unit boundaries within which cases are aggregated, a problem referred to in the spatial context as the “modifiable areal unit problem” [26] (an analogous, but less discussed in the literature, issue arises with temporal units).

Importantly, to avoid errors due to ecologic fallacy, hypotheses should be conceptualized and analyzed at the same level of aggregation—that is, if an exposure of interest is at the individual level (e.g., marijuana use among drivers as a cause of motor vehicle fatalities [10]), it should be analyzed and interpreted at the individual level, whereas when an exposure of interest is at the group level (e.g., marijuana decriminalization as a cause of change motor vehicle fatality rates [27]), it should be analyzed and interpreted at the group level. This fallacy may occur at the conceptual stage of a project—group-level factors such as enacted policies that do not confound at the individual-level may confound group-level associations between exposures and outcomes while individual-level characteristics that confound associations between individual exposure and outcomes may not have an analogous exposure confounding group-level exposures and outcomes [28]. Note also that measurement artifacts can impact ecologic studies in ways not familiar to researchers used to individual-level studies—when individual-level data are aggregated up to group-level metrics, choices made in expressing aggregated variables as proportions (e.g., percent of people living in poverty) or continuums (e.g., per capita income) may strongly affect expected directions of bias even in the presence of non-differential measurement error [29, 30].

Ecological designs in injury research frequently assess the impacts of policies. For example, Branas and Knudsen used a cross-sectional ecologic design to assess the association of motorcyclist helmet laws with motorcyclist death rates in FARS [6]. Mooney et al. used a cohort design to estimate that state-level Complete Streets policy implementation was associated with an increase in commuter cyclists and a decrease in cyclist fatality rates using data from FARS [8], and Aydelotte et al. used a difference-in-differences cohort design to examine the impact of recreational marijuana legalization on motor vehicle fatalities, also using FARS data [31].



## Case-control

In contrast to the ecologic design, in a case-control design registry data are used to identify individual cases to which controls sampled from another dataset or an underlying population are matched. This approach allows for straight-forward analysis of causes of the injury event itself—that is, under the assumption that the controls represent the same underlying population that cases arose from, exposures that are more prevalent among cases than controls, after adjustment for confounding factors, may be causes of the injury event itself. For case-control studies to be correctly designed and interpreted, it must be recognized that the case series is generated from an underlying cohort and the controls are sampled from the cohort that generated the cases to estimate the prevalence of exposure in this source population [32, 33].

A feature that is perhaps unique to injury registry data is that for case-control studies the investigator can consider each injury occurrence from one of several units of analysis, including the injured person, the location of the injury, or the at-fault party. The decision on what unit of analysis to use in the design affects the hypotheses that can be tested and the variables that can be used in the analyses [34]. Consider the following three hypothetical case-control studies of pedestrian injury risk drawing from FARS: a person-based case-control study, a location-based case-control study, and a responsibility or culpability study.

In the first study, a series of motor vehicle fatality entries in FARS could be analyzed in a person-level case-control design. In this study, matched control drivers would be recruited to provide data on their personal characteristics (e.g., age and sex) and behaviors (e.g., were they driving at the same time of day as the case driver, were they under the influence of alcohol at that time [35]?). In this design, because all variables can be conceptualized and measured for both cases and controls, variables related to individuals, like age, sex, driving while under the influence of alcohol, could be analyzed as exposures (predictors), confounders, mediators, or effect modifiers. An analysis of etiological heterogeneity could be conducted using the same data by classifying cases using variables that describe inherent features of collision, such as whether the injured party died or was admitted to the hospital. In this analysis, each sub-type of cases would be compared to its matched controls and the extent to which the sub-type specific odds ratios differ is a measure of etiological heterogeneity [36].

In the second study, the same case series of motor vehicle fatality events would be selected from FARS, but rather than being matched by people who could have

been killed but were not, they would be matched to places where fatalities could have occurred but did not. At all sampled locations, characteristics of street segments and intersections would be assessed. Then, characteristics of the location (e.g., presence or absence of traffic calming infrastructure or an alcohol selling establishment) can be used as exposure variables and tested for associations with case vs. control locations, contributing information about the environmental risk factors potentially contributing to the fatality [37]. Furthermore, the case locations could be categorized by circumstances of the crash, such as the victim's gender or age or the driver's sobriety, allowing for an etiologic heterogeneity design. However, by contrast to the person-based case-control design, in the location-based design, control locations cannot be categorized in this manner because the crashes leading to fatalities have not occurred at control locations. Thus, in a location-based case-control design, characteristics of the driver can be used to design a study of etiological heterogeneity—are characteristics of the location associated with different types of injuries. However, variables related to the driver or crash circumstances cannot be used as measures of exposures, confounders, mediators, or effect modifiers [36, 38]. Note that in the location-based case-control design, characteristics of the crash *location* could be considered as exposures, confounders, mediators, or effect modifiers, which they could not in the person-based case-control design.

In the third study, the cases are drivers deemed responsible for the crash and controls are drivers not responsible for crashes. A subtype of this design, sometimes called a quasi-induced exposure design, matches drivers involved in the same 2-vehicle crash. The underlying logic of this design is that the drivers involved in but not responsible for the crash serve as controls (matched controls in the case of the quasi-induced exposure design) that can be used to estimate the underlying prevalence of exposures or characteristics of the population of non-culpable drivers. This assumes that non-culpable drivers involved in a crash are a random sample of drivers (in the quasi-induced exposure, a random sample conditional on the matching factors—time and place of driving). As in the person-based case-control design, characteristics of drivers such as age or intoxication could be assessed and analyzed as exposures, confounders, mediators, or effect modifiers. However, as compared with the conventional person-based case-control design, these variables would be predictors of being *responsible* for a collision, not for being in a collision at all, which is a subtly different outcome for two key reasons: first, any variables included in the responsibility assessment procedure cannot be analyzed (e.g., if intoxication is considered when deciding which driver is responsible for a collision). Second, binary responsibility assessment is an inherently challenging process and likely

includes some error (e.g., if driver A made a risky move that driver B could have avoided had driver B been paying better attention, does driver B still represent a random sample of the driving population) which may bias results [39].

Advantages of the case–control approach include theoretical rigor with which relates case–control designs to underlying cohort designs and the ability to directly assess factors contributing to injury risk. The key disadvantage of this approach is the challenge of identifying a dataset containing controls that truly represent the same source population as the cases and for whom similarly specified variables are available. Accordingly, the key issue with this design is accounting for differences between the cases and controls, both in sampling processes leading to incorporation in the dataset and in variable specification.

For example, both Li et al. [9] and Romano et al. [40] compared drug and alcohol consumption in motor vehicle collision fatality cases to drug and alcohol consumption in a control group selected from drivers agreeing to roadside testing. Under the assumption that the controls represent the population that gave rise to the collision set, the greater prevalence of drug and alcohol use identified among cases suggests that drugs and alcohol contribute to motor vehicle fatalities. However, if people who have used drugs or alcohol systematically refuse participation in the roadside study, these results overestimate the elevation of risk due to drug and alcohol use.

### Case-only (Sometimes Called Case series)

A case series design is similar to a case–control design, except that the comparison group is selected from within the registry and case types are compared to one another. Associations are estimated at the individual or location level in relation to some implicit or explicit causal hypothesis. In this design, the case series is sub-typed into two or more groups by some variable (e.g., age of the injured party) and then this variable becomes the dependent variable in the case-only analyses [9, 10]. Case-only design results are only interpretable if the case series is understood to have been conceptually generated from a cohort that otherwise would have been analyzed using cohort or case–control methods. That is, the case series in a case-only design is the same case series that otherwise would have been analyzed in a case–control study. There are two primary flavors of case-only design—in the “etiologic heterogeneity” design, cases are categorized by some aspect of case status that has no analogous interpretation in controls. In the “interaction” design, cases are categorized by an exposure that could be measured in a control and under certain assumptions the case-only analysis provides an estimate of multiplicative interaction effects. For example, a case-only study of suicide using NVDRS to explore

decedent age and means (e.g., comparing firearm suicides to all other suicides) is an etiologic heterogeneity design, because cases are classified by a variable, means of suicide, that is a feature of the case with no analogous construct for controls. A case-only study of suicide using NVDRS that stratified cases into age groups and firearm sales volume within the state in which the decedent lived would be a test of multiplicative interaction, because both age and firearm volume sales at the state level are interpretable for people who would be eligible to be controls in a case–control analysis of the case series.

The primary advantage of the case series approach is the ease of conducting such a study—the data are available and the analytic techniques are simple. However, there are substantial disadvantages, including the implausibility of assumptions (for interaction designs) and limited utility of the scope of inquiry within injury (for etiologic heterogeneity designs). In case-only analyses, the statistical parameter resulting from an analysis can only be interpreted when the case series is understood to have been generated from an underlying cohort, and defining that cohort may be challenging if registry capture is incomplete [32, 41, 42]. Finally, because the distinction between etiologic heterogeneity and interaction designs is not widely appreciated, analysts may incorporate variables inappropriate for the target of inference. For example, in a case-only analysis comparing male versus female decedents with the intent of exploring interactions between sex and gun ownership, an analyst might incorporate adjustment variables such as the cause of death, which cannot be conceptualized for a comparable control. When such variables are included in a regression model, the scientific question is obscured and the covariate-adjusted effect estimates are difficult to interpret.

Thus, the key challenge in the case-only design is identifying a clear scientific question for which this analogous cohort is identifiable and the assumptions necessary to interpret the results hold. Like all causally focused analyses, case-only designs require the researcher to choose a counterfactual within a unit of analysis—what the researcher imagines could be changed to prevent the injury or improve injured parties’ outcomes. This choice is central to the analysis—it impacts the conceptualization of the underlying at-risk population, the comparison of interest, whether selected variables should be considered confounders, mediators, or effect modifiers, and the interpretation of any estimated effects. Yet in case-only designs, it is not always clearly stated how the analytic comparison relates to the underlying question, largely because the distinction between the etiologic heterogeneity design and the interaction design is not well understood.

For example, Kaplan et al. explored determinants of firearm suicide among adults using data from NVDRS [43]. The primary results from this analysis determined that, among

both men and women, age and veteran status were associated with firearm suicide as compared with suicide among other means. This result can be understood as an etiologic heterogeneity finding—age and veteran status are associated with means of suicide, which is a variable that can only be used to distinguish sub-groups of cases and would not be applicable to controls—but cannot say anything about suicide prevention overall.

Finally, a less common flavor of case series design involves comparing cases' exposure to a transient exposure at the time of an index injury to exposure level in that same subject at another time. This design, sometimes called an individual-level case-crossover design, estimates the temporary risk elevation associated with that exposure. This design is appealing for its simplicity and because it accounts for time-fixed individual-level confounding. However, it requires exposure assessment at a time where the injured subject was not injured, which is uncommon in registries whose focus is to record injuries. In cases where injury registry can be linked to external data sources (e.g., when Finnish occupational injury registry data were linked to payroll records to assess risks of working selected hours [44]), this design is appealing.

## Outcomes Research

Some registries (e.g., NTDB) include records of care and follow-up after the injury and others can be linked to such outcome data such as medical records, arrest records, and death certificates. These datasets, including an injury event and its outcomes, can then be used to research the consequences of injury events and to identify potentially modifiable environmental or clinical conditions that affect injury outcomes. In this design, the registry serves to define a cohort or sampling frame, typically considering the event causing the individual to join the cohort as baseline and following up through linked data or care records.

In some cases, electronic registries have served as an efficient platform for recruitment, randomization, and follow-up for pragmatic randomized clinical trials (e.g., [45]), though this approach has not been widely adopted within injury outcomes research [46]. This is likely because the integration needed to ensure electronic health record systems report to injury registries in real-time might come at the cost of data quality monitoring, which is already a concern for registries [47]. Nonetheless, as automated approaches to identifying and flagging errors proliferate [48], this approach may offer exciting opportunities for registry-based injury outcomes randomized trials.

Advantages of registry-based outcomes research include data availability—even registries designed to track incidence provide rich baseline characterization—and wide population

coverage. Disadvantages, as compared with hospital-based outcome research relying on the full medical record, include the limitation that only data abstracted into the registry or linkable data is available, limiting investigators to variables selected by the registry for harmonization across sites.

Key issues for this design include challenges around linking registry records to external datasets—because registrants are typically not asked to consent to being included in a registry, access to personal identifiers used to link to external data are rightly limited—and record linkage software can be challenging to implement and may induce selection biases due to incomplete linkage.

For example, Sato et al. used records from the Victoria State Trauma Registry, a trauma registry that routinely links medical records of major trauma patients in all care facilities in Victoria, Australia, to death records in the same state, to examine in-hospital mortality and other outcomes among older adult patients who had undergone major trauma [49].

## Conclusions

Injury events are rare and are frequently captured in registries. Different design choices in analysis of these registries' data affect the results' interpretation. The key first step for a researcher is to choose which counterfactual (if any) within which unit of analysis is of interest—that is, what does the researcher imagine could be changed and at what level of organization (e.g., person, neighborhood, and state) to prevent injuries or improve injured parties' outcomes. Working from this hypothetical counterfactual, units might be individual people (e.g., when studying characteristics of the injured party or an at-fault party) interventions on individual people (e.g., when studying treatments received in post-injury care) or individual places (e.g., when studying the physical environment at the location of the injury event). Analytic units could also be groups of people or places, (e.g., when studying states included in an ecological cohort). The choice of counterfactual and unit of analysis is fundamental to the scientific process, impacting the conceptualization of the underlying at-risk population, the comparison of interest, whether selected variables should be considered confounders, mediators or effect modifiers, and the interpretation of any estimated effects. There are examples of analyses of registries in the literature where the analyzed data are drawn from multiple units of analysis—characteristics of the injured party, the location, and the at-fault party—which may or may not be measurable among controls or the underlying cohort. Because the applicable unit of analysis and its relationship to the underlying population of such units is obscured, the results of the analyses are not readily interpretable.



In summary, registry data can be analyzed using an assortment of study designs, each with their strengths and drawbacks, and it is important that investigators and consumers of the research results understand the strengths and drawbacks of each.

**Funding** This work was supported by grant R00LM012868 from the National Library of Medicine, grants R01AA028552 and K01AA026327 from the National Institute on Alcohol Abuse and Alcoholism, and grants R49-CE003094 and R49CE003087 from the Centers for Disease Control and Prevention.

## Declarations

**Conflict of Interest** The authors declare no competing interests.

## References

- Haddon W, Bradess VA. Alcohol in the single vehicle fatal accident: experience of Westchester County, New York. *J Am Med Assoc.* 1959;169(14):1587–93.
- Sleet DA, Baldwin G, Marr A, et al. History of injury and violence as public health problems and emergence of the National Center for Injury Prevention and Control at CDC. *J Safety Res.* 2012;43(4):233–47.
- Porta M. *A dictionary of epidemiology.* Oxford university press; 2014.
- National Highway Transport Safety Administration. MMUCC guideline: Model Minimum Uniform Crash Criteria. Accessed January 28, 2022. [https://www.ghsa.org/sites/default/files/publications/files/MMUCC\\_5thEd\\_web.pdf](https://www.ghsa.org/sites/default/files/publications/files/MMUCC_5thEd_web.pdf)
- National Center for Statistics and Analysis. Traffic safety facts: pedestrians.; 2019. Accessed January 27, 2022. <https://crashstats.nhtsa.dot.gov/Api/Public/Publication/813079>
- Branas CC, Knudson MM. Helmet laws and motorcycle rider death rates. *Accid Anal Prev.* 2001;33(5):641–8.
- Gorman DM, Huber Jr JC, Carozza SE. Evaluation of the Texas 0.08 BAC law. *Alcohol Alcohol.* 2006;41(2):193–199.
- Mooney SJ, Magee C, Dang K, et al. “Complete Streets” and adult bicyclist fatalities: applying G-computation to evaluate an intervention that affects the size of a population at risk. *Am J Epidemiol.* 2018;187(9):2038–45.
- Li G, Brady JE, Chen Q. Drug use and fatal motor vehicle crashes: a case-control study. *Accid Anal Prev.* 2013;60:205–10.
- Li G, Chihuri S, Brady JE. Role of alcohol and marijuana use in the initiation of fatal two-vehicle crashes. *Ann Epidemiol.* 2017;27(5):342–347. e1.
- McCartt AT, Northrup VS. Factors related to seat belt use among fatally injured teenage drivers. *J Safety Res.* 2004;35(1):29–38.
- Amoros E, Chiron M, Thélot B, Laumon B. The injury epidemiology of cyclists based on a road trauma registry. *BMC Public Health.* 2011;11(1):1–12.
- Elvik R. Comparative analysis of techniques for identifying locations of hazardous roads. *Transp Res Rec.* 2008;2083(1):72–5.
- Tian N, Cui W, Zack M, Kobau R, Fowler KA, Hesdorffer DC. Suicide among people with epilepsy: a population-based analysis of data from the US National Violent Death Reporting System, 17 states, 2003–2011. *Epilepsy Behav.* 2016;61:210–7.
- Hemenway D, Solnick SJ. The epidemiology of homicide perpetration by children. *Inj Epidemiol.* 2017;4(1):1–6.
- Athey S, Imbens GW. Identification and inference in nonlinear difference-in-differences models. *Econometrica.* 2006;74(2):431–97.
- Donald SG, Lang K. Inference with difference-in-differences and other panel data. *Rev Econ Stat.* 2007;89(2):221–33.
- Goodman-Bacon A. Difference-in-differences with variation in treatment timing. *J Econom.* 2021;225(2):254–77.
- Schwartz S. The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. *Am J Public Health.* 1994;84(5):819–24.
- Ward JW, Michalek JJ, Azevedo IL, Samaras C, Ferreira P. Effects of on-demand ridesourcing on vehicle ownership, fuel consumption, vehicle miles traveled, and emissions per capita in US States. *Transp Res Part C Emerg Technol.* 2019;108:289–301.
- Branas CC, Cheney RA, MacDonald JM, Tam VW, Jackson TD, Ten Have TR. A difference-in-differences analysis of health, safety, and greening vacant urban space. *Am J Epidemiol.* 2011;174(11):1296–306.
- Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: estimating the effect of California’s tobacco control program. *J Am Stat Assoc.* 2010;105(490):493–505.
- Robinson WS. Ecological correlations and the behavior of individuals. *Am Sociol Rev.* 1950;15(3):351–7.
- Piantadosi S, Byar DP, Green SB. The ecological fallacy. *Am J Epidemiol.* 1988;127(5):893–904.
- Selvin HC. Durkheim’s suicide and problems of empirical research. *Am J Sociol.* 1958;63(6):607–19.
- Fotheringham AS, Wong DW. The modifiable areal unit problem in multivariate statistical analysis. *Environ Plan A.* 1991;23(7):1025–44.
- Mark Anderson D, Hansen B, Rees DI. Medical marijuana laws, traffic fatalities, and alcohol consumption. *J Law Econ.* 2013;56(2):333–69.
- Morgenstern H. Ecologic studies in epidemiology: concepts, principles, and methods. *Annu Rev Public Health.* 1995;16(1):61–81.
- Brenner H, Savitz DA, Jöckel KH, Greenland S. Effects of non-differential exposure misclassification in ecologic studies. *Am J Epidemiol.* 1992;135(1):85–95.
- Mooney SJ, Richards CA, Rundle AG. There goes the neighborhood effect: bias due to non-differential measurement error in the construction of neighborhood contextual measures. *Epidemiol Camb Mass.* 2014;25(4):528.
- Aydelotte JD, Brown LH, Luftman KM, et al. Crash fatality rates after recreational marijuana legalization in Washington and Colorado. *Am J Public Health.* 2017;107(8):1329–31.
- Wacholder S. Design issues in case-control studies. *Stat Methods Med Res.* 1995;4(4):293–309.
- Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies: I. Principles. *Am J Epidemiol.* 1992;135(9):1019–28.
- Kim JH, Mooney SJ. The epidemiologic principles underlying traffic safety study designs. *Int J Epidemiol.* 2016;45(5):1668–75.
- Branas CC, Richmond TS, Ten Have TR, Wiebe DJ. Acute alcohol consumption, alcohol outlets, and gun suicide. *Subst Use Misuse.* 2011;46(13):1592–603.
- Begg CB, Zhang ZF. Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiol Prev Biomark.* 1994;3(2):173–5.
- Mooney SJ, DiMaggio CJ, Lovasi GS, et al. Use of Google Street View to assess environmental contributions to pedestrian injury. *Am J Public Health.* 2016;106(3):462–9.
- Krumkamp R, Reintjes R, Dirksen-Fischer M. Case–case study of a Salmonella outbreak: an epidemiologic method to analyse surveillance data. *Int J Hyg Environ Health.* 2008;211(1–2):163–7.
- Brubacher J, Chan H, Asbridge M. Culpability analysis is still a valuable technique. *Int J Epidemiol.* 2014;43(1):270–2.

40. Romano E, Torres-Saavedra P, Voas RB, Lacey JH. Marijuana and the risk of fatal car crashes: what can we learn from FARS and NRS data? *J Prim Prev.* 2017;38(3):315–28.
41. Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene environment interaction: case-control studies with no controls! *Am J Epidemiol.* 1996;144(3):207–13.
42. Gatto NM, Campbell UB, Rundle AG, Ahsan H. Further development of the case-only design for assessing gene–environment interaction: evaluation of and adjustment for bias. *Int J Epidemiol.* 2004;33(5):1014–24.
43. Kaplan MS, McFarland BH, Huguet N. Characteristics of adult male and female firearm suicide decedents: findings from the National Violent Death Reporting System. *Inj Prev.* 2009;15(5):322–7.
44. Härmä M, Koskinen A, Sallinen M, Kubo T, Ropponen A, Lombardi DA. Characteristics of working hours and the risk of occupational injuries among hospital employees: a case-crossover study. *Scand J Work Environ Health.* 2020;46(6):570.
45. Fröbert O, Lagerqvist B, Olivecrona GK, et al. Thrombus aspiration during ST-segment elevation myocardial infarction. *N Engl J Med.* 2013;369:1587–97.
46. Johnson EA, Carrington JM. Clinical research integration within the electronic health record: a literature review. *CIN Comput Inform Nurs.* 2021;39(3):129–35.
47. Rubinger L, Ekhtiari S, Gazendam A, Bhandari M. Registries: big data, bigger problems? *Injury:* Published online; 2021.
48. Zhang JM, Harman M, Ma L, Liu Y. Machine learning testing: survey, landscapes and horizons. *IEEE Trans Softw Eng:* Published online; 2020.
49. Sato N, Cameron P, Mclellan S, Beck B, Gabbe B. Association between anticoagulants and mortality and functional outcomes in older patients with major trauma. *Emerg Med J:* Published online; 2021.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.