

A concise guide to reproducible research using secondary data

Maryna Ivets, Eva Goetjes, Kai Miele, Katharina Blankart

2021-09-28

A study guide to empirical research in economics and management using secondary data

Objective

The objective of this study guide is to provide a resource to graduate students, early career PhD candidates and researchers performing applied empirical research in economics and management sciences. The practical application is in the field of analysis of health care markets using secondary data. Many textbook examples use readily available data sets for analysis of econometric problems. For students aiming to conduct research by generating their own analysis data set, important steps that lead to a final analysis data set are missing. Besides, many resources focus on labour economics problems. Resources that consider processing and generating secondary data are scarce. One reason is that often these data sources are subject to confidentiality and data protection issues such that these cannot be made available.

The objective of this study guide is to provide a resource that covers the major steps of a reproducible research project in 5 steps. We will introduce important terminology, highlight the relevant tasks to be performed, and to provide key resources in the form of text books and websites available via open access. We aim to provide a concise guide that users of this guide can easily access when starting academic research. Each section is to be read within 10-15 minutes. For this reason, we will not cover any specific data science or econometric method, but point to the relevant resources. Therefore, users of this guide should have basic knowledge in statistics, econometrics and program evaluation. We further recommend basic knowledge in a statistical package such as R or Stata. For the study guide to be of most benefit, readers should have background knowledge and a research idea in mind while reading.

Learning objectives

The goal is to set up and carry out a data science project using secondary data. Students will learn all steps starting with hypothesis formulation, data generation and analysis, and presentation of empirical results.

After reading and applying the principles introduced in this study guide, you will be able to:

1. Recognize the features of using secondary (health care) data in empirical research.
2. Execute the steps of a reproducible research project.
3. Be able to implement an empirical research project.
4. Recall the steps taken to execute a reproducible research project using secondary data.

Structure of the study guide

The study guide consists of five chapters that include the essential steps of a reproducible research project. Each step of reproducible research is covered in three parts.

1. An introduction to the basics concepts and key terminology.
2. A resources box that includes reference material how to perform this step including main textbooks and references to current web resources. We will emphasize open source materials.
3. A showcase example of an empirical project reproduced based on the article: *Hellerstein, Judith K. 1998. "The Importance of the Physician in the Generic versus Trade-Name Prescription Decision." The RAND Journal of Economics 29 (1): 108–36. <https://doi.org/10.2307/2555818>.*

This is a living document

How you can contribute to this study guide? Best practices how to perform reproducible research are constantly developing. We aim to keep resources up to date. If you come across a good open access resource or suggestions for improvement, please share it by emailing to:

Acknowledgements

The authors thank Christoph Kronenberg for comments and suggestions.

Development of this resource has received funding by *Data Literacy Education.nrw*. It is part of the DataCampus project of the University of Duisburg-Essen.

Introduction to reproducible research

“Only results that can be replicated are truly scientific results. If there is no chance to replicate research results, they can be regarded as no more than personal views in the opinion or review section of a daily newspaper.” (Huschka 2013)

What is reproducible research?

Scientific journal editors and research funders are increasingly promoting transparency in research. To encourage the principles of reproducible research, the related institutions are requesting authors to make their research reproducible. To overcome criticisms regarding the validity and power of empirical tests, this means that data and program code need to be shared upon manuscript acceptance, or earlier stages of the submission process. The purpose is that third-parties have the possibility to reproduce the content and analysis of a study, and conclusions of a study on their own.

Efforts to increase reproducibility have been expressed by multiple institutions within the social sciences, for example:

- The best practices statement by the Social Science Data Editors,
- In Germany, by the German Research Foundation (DFG) and the Consortium for the Social, Behavioural, Educational and Economic Sciences RatWSD
- In editorial statement of journals, for example American Economic Review, Management Science or the Journal of International Business Studies (Meyer, Witteloostuijn, and Beugelsdijk 2017; Orozco et al. 2020)

Generally, **reproducibility** of research can be defined as “the ability of a researcher to duplicate the results of a prior study using the same materials and

procedures as were used by the original investigator. In an attempt to reproduce a published statistical analysis, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis to determine whether they yield the same results.” (Bollen et al. 2015). Study results are considered reproducible if after an article publication another researcher can conduct the analyses using identical data and obtain the same results using the material provided.

Since empirical economic research is based on the application of a code to a dataset to answer a pre-defined research question, ensuring the reproducibility includes sharing the data and code to allow others to re-analyze the data and to reproduce the reported results (Orozco et al. 2020). To achieve this, the data and code need to be properly managed while working on a project.

Another concept closely related to *reproducibility* is the concept of *replicability* that “refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected.” (Bollen et al. 2015). Thus, for example, if an investigator tries to replicate a scientific finding that documents a relationship between two or more variables by using the same scientific methodology but in a new setting, i.e. with new data, and fails and to reach a similar conclusion, i.e. replicate it – a failure to replicate occurs. The opposite is said to be true if the results are replicated. Thus, *reproducibility* and *replicability* are considered to be the two main elements of empirical research.

To adapt reproducible practices early on, undergraduate and graduate students are encouraged to perform reproducible research in their term papers, theses, and practical applications as soon as possible. For this reason, university teachers are increasingly asking to submit reproduction material (source data information, data programming and analysis code) at all stages of study.

The concept of reproducible research is not new and goes back as far as the late 1800s (Vilhuber 2020). However, reproducibility studies have disclosed that many researchers do not follow the principles of reproducible research. At the same time, the increase in availability and use of public and especially non-public data sources, and the increase in the reliance of research methods based on specific software bringing the principles of reproducible research on the agenda of many researchers.

The resources box in each chapter provides material of best practices in data and code management to perform a reproducible research project, as well as provides resources for best practices in cleaning the data and conducting data analysis.

Why you, as a student and researcher, should care about reproducible research?

Avoid common biases that lead to biased or false research results

There is a number of threats to reproducible research process that can undermine scientific research or lead to false or biased conclusions and publications, see, e.g. @ref(fig:munafo) illustrating main threats to reproducible science (Munafo et al. 2017).

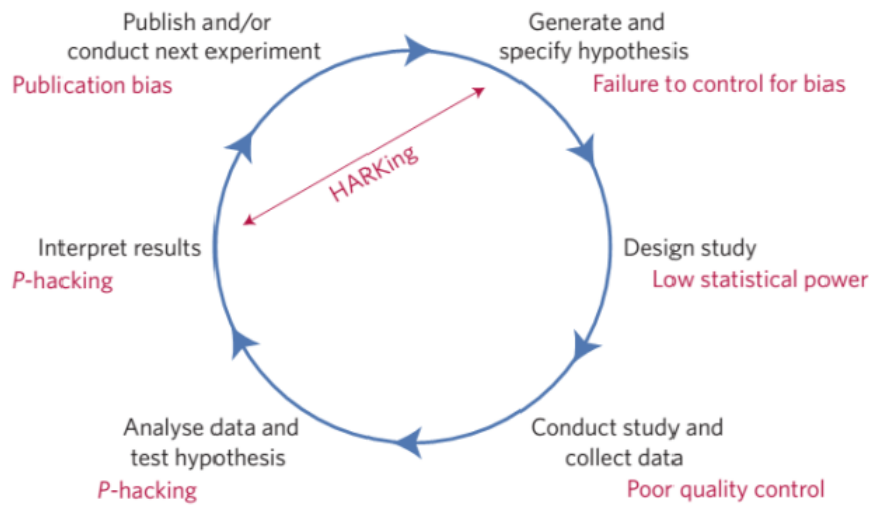


Figure 1: munafo

Increase productivity of your work and the work of the scientific community by performing reproducible research

Following the reproducible research principles allows you to be a part of a good academic practice that strives to improve the quality, efficiency and reliability of scientific research.

Main steps to produce reproducible research

To follow the best practices of reproducible research, you will need to consider *how* to perform your project and the *steps a reproducible research project* contains.

There are three main principles to enhance reproducible research (Orozco et al. 2020):

1. ***Organize your work:*** consider and plan your steps at the beginning of the project
2. ***Code for others:*** set up each step of your project such that an outsider could follow your documentation
3. ***Automate as much as you can:*** avoid processing analyses and results using point-and-click software (MS Excel), export results directly and create a reproducible project documentation.

In this study guide we follow the main steps of performing a reproducible data science project (Bezjak et al. 2018). To make your empirical research study process reproducible, you need to follow these five steps (Figure @ref(fig:reproducible)):

1. ***Formulating a hypothesis*** (Section @ref(hypothesis))
2. ***Designing the study*** (Section @ref(designstudy))
3. ***Running the study and collecting the data*** (Section @ref(run-study))
4. ***Analyzing the data*** (Section @ref(analyzing))
5. ***Reporting the study*** (Section @ref(report-study))

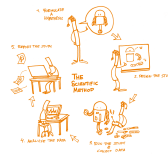


Figure 2: reproducible

Secondary (health care) data

This study guide concentrates on empirical investigation using secondary data, that means data that is not originally collected or generated by the researcher for the purpose of the study. Secondary data covers any existing data generated by companies, institutions, and individuals.

Examples of secondary data sources are routinely collected health data (administrative claims, electronic medical records), bibliometric data, survey data, regulatory data, data generated in mobile applications.

Creating an environment for productive research projects

Before you start running the study you should assess your available resources, that means time and own expertise to see whether the study design is feasible. If there is an opportunity to work collaboratively and you think that the project can benefit from the expertise of another researcher, you can approach the person and ask whether they would be interested to work together on the project. Joining resources might relax your resource availability constraints. It may further ensure that you engage in a reproducible research process.

Resources Box

- An overview article of the principles of reproducible research and practical application: Orozco V, Bontemps C, Maigné E, Piguet V, Hofstetter A, Lacroix A, et al. How to Make a Pie: Reproducible Research for Empirical Economics and Econometrics. *Journal of Economic Surveys*. 2020;34(5):1134–69. <https://doi.org/10.1111/joes.12389>
- A guideline to minimize the risk of reporting false positives (type I errors), improve the quality of hypothesis-testing research and statistical reporting: Meyer, Klaus E., Arjen van Witteloostuijn, and Sjoerd Beugelsdijk. 2017. “What’s in a p? Reassessing Best Practices for Conducting and Reporting Hypothesis-Testing Research.” *Journal of International Business Studies* 48 (5): 535–51. <https://doi.org/10.1057/s41267-017-0078-8>.
- The Worldbank blog provides “A Curated List of Our Postings on Technical Topics – Your One-Stop Shop for Methodology”

Formulating a hypothesis



Figure 3: idea

Source: DLPNG

Basic steps to formulate a hypothesis

You decided to undertake a scientific project. Where do you start? First, you need to find a research question that interests you and formulate a hypothesis. We will introduce some key terminology, steps you can undertake and examples how to develop research questions. Note that there is no single best way to develop a research idea (Pischke 2012).

What if someone assigns a topic to me? For students attending undergraduate and graduate courses that often pick topics from a list, all of these steps are equally important and necessary. You still need to formulate a research question and a hypothesis. And it is important to clarify the relevance of your topic for yourself.

What is a hypothesis?

A hypothesis is a statement that introduces your research question and suggests the results you might find in your research. You will test your hypothesis with your empirical analysis. Thus, hypothesis constitutes the main basis of your scientific investigation. You should be careful when creating it.

A hypothesis is an educated guess. You start by posing an economic question and formulate a hypothesis about this question. Then you test it with your data and analyses and either accept or reject the hypothesis.

How do you develop a research question and formulate a hypothesis?

When thinking about a research question (Trochim n.d.), you need to identify a topic that is

- *Relevant*, important in the world and interesting to you as researcher: Does working on the topic excites you? You will spend many hours working on it. Therefore, it should be interesting and engaging enough for you to motivate your continued work on this topic.
- *Specific*: not too broad and not too narrow
- *Feasible* to research within a given timeframe: Is it possible to answer the research question based on your time budget, data and additional resources.

How do you find a topic or develop a feasible research idea in the first place? Finding an idea is not difficult, the critical part is to find a **good** idea. How do you do that? There is no one specific way how one gets an idea, rather there is a myriad of ways how people come up with potential ideas (Varian 2016).

Once you decide on an idea, run it by a few people. That means approach your supervisor and other fellow students to see whether it is interesting enough to pursue it.

For example, you can find inspiration by

- Looking at insights from the world around you: your own life and experiences, observe the behavior of people around you
- Talking to people around you, experts, other students, family members
- Talking to individuals outside your field (non-economists)
- Talking to professionals working in the area you are investigating (you may use social media and professional platforms like LinkedIn or Twitter to make contact)
- Reading journal articles from other non-economic social sciences and the medical literature
- Reading economic and non-economic newspaper and magazine articles (e.g. Süddeutsche, Frankfurter Allgemeine, Tagesspiegel, The Economist, the Wall Street Journal, The New York Times, The Guardian), watching TV programs
 - What are the issues being discussed?
 - How do these issues affect people's lives?

You could in addition

- Go to virtual and in-person seminars, for example the Essen Health Economics Seminar
- Look at abstracts of scientific articles and working papers
- Look at the literature in a specific field you are interested in, for example screening complete issues of journals or editorials about certain research advancements. By reading this literature you might come up with the idea on how to extend and refine previous research.

Develop a Hypothesis

Before you formulate your hypothesis, read up on the topic of interest. This reading should provide you with sufficient information to narrow down your research question. The research question comes from the topic you are contemplating. Once you find your question you need to develop a hypothesis. A hypothesis is your research question distilled into a one sentence statement. It presents a statement of your expectations regarding your research question's results. You propose to prove your hypothesis with your research by testing the relationship between two variables of interest. Thus, a hypothesis should be testable with the data at hand. There are two types of hypotheses: alternative or null. Null states that there is no effect. Alternative states that there is an effect.

It should be noted that there is an alternative view on this that suggests one should not look at the literature too early on in the idea-generating process in order not to be influenced and shaped by someone else's ideas. According to this view you can spend some time (i.e. a few weeks) trying to develop your own original idea. Even if you end up with an idea that has already been pursued by someone else, this will still provide you with good practice in developing publishable ideas. After you have developed an idea and made sure that it was not yet investigated in the literature, you can start conducting a systematic literature review. By doing this, you can find some other interesting insights from the work of others that you can synthesize in your own work to produce something novel and original.

Identify Relevant Literature

For your research project you will need to identify and collect previous relevant literature. It should involve a thorough search of the keywords in relevant databases and journals. Place emphasis on articles from high-ranking journals with significant numbers of citations. This will give you an indication of the most influential and important work in the field. Once you identify and collect the relevant literature for your topic, you will need to **critically** synthesize it in your literature review.

When you perform your literature review, consider theories that may inform your research question. For example when studying physician behavior, you may consider approaches made using principal-agency theory.

Research question or literature review, a hen and egg problem?

Whether you start reading the literature first or by developing an idea may depend on your stage (graduate student, early career researcher) and other goals. However, thinking freely about what you like to investigate first may help to critically develop a feasible and interesting research question.

We highlight an example how to start with investigating the real-world and subsequently posing a research question ("How to Write a Strong Hypothesis Steps and Examples" 2019; "Developing Strong Research Questions Criteria and Examples" 2019; Schilbach 2019). For example, based on your own observation you notice that people seem to spend extensive amount of time looking at their smartphones. Maybe even you yourself engage in the same behavior. In addition, you read a BBC News article Social media damages teenagers' mental health, report says.

\begin{figure}



{

}

\caption{social_media} \end{figure}

Source: BBC

You decided to translate this article and your observations into a research question. A **research question** could be: *How does daily social media use influence our mental health?* Before you formulate your hypothesis, read up on the topic of interest. This reading should provide you with sufficient information to narrow down your research question. Read economic, medical and other social science literature on the topic. There is likely to be a vast amount of non-econ literature from non-econ fields that are doing research on your topic of interest, for example psychology or neuroscience. Familiarize yourself with it and master it. Do not get distracted by different scientific methodologies and techniques that might seem not up-to-par to the economic studies (small sample sizes, endogeneity, uncovering association rather than causation, etc.), but rather focus on suggestions of potential mechanisms.

A hypothesis is then your research question distilled into a one sentence statement. It presents a statement of your expectations regarding your research question's results. You propose to prove your hypothesis with your

research by testing the relationship between two variables of interest. Thus, a hypothesis should be testable with the data at hand. There are two types of hypotheses: alternative or null. The **null*** hypothesis states that there is no effect. The **alternative** hypothesis states that there is an effect.

A hypothesis related to the above-stated research question could be: *The increased use of social media among teenagers leads to (is associated with) worse mental health outcomes, i.e. increased incidence of depression, worse well-being and lower self-esteem.* Your hypothesis suggests a direction of a relationship that you expect to find. It is not a blind guess, but is guided by your observations and existing evidence, that means the reports that you have collected. It is testable with scientific research methods, that means by using statistical analysis of the relevant data.

Your hypothesis suggests a relationship between two variables: social media use (your independent variable *X*) and mental health (dependent variable *Y*).

Note, that this hypothesis could be framed in terms of correlation (is associated with) or causation (leads to). This should be reflected in the choice of scientific investigation you decide to undertake.

The **null** hypothesis is: *There is no relationship between social media use among teenagers and their mental health.*

It should be noted that there is an alternative view on this idea - a hypothesis-generating process that suggests that one should not look at the literature too early on in the idea-generating process not to be influenced and shaped by someone else's ideas. According to this view you can spend some time like a few weeks trying to develop your own original idea. Even if you end up with an idea that has already been pursued by someone else and published, this will still provide you with good practice in developing publishable ideas.

After you have developed an idea and made sure that it was not yet investigated in the literature, you can start conducting a systematic literature review. By doing this, you can find some other interesting insights from the work of others that you can synthesize in your own work to produce something novel and original.

Resources box

How to develop strong research questions

- The form of the research process

Identify relevant literature from major general interest and field literature

To identify the relevant literature you can

- use academic search engines such as Google Scholar, Web of Science, Econ-Lit, PubMed.
- search working paper series such as the National Bureau of Economic Research, NetEc or IZA
- search more general resource sites such as Resources for Economists
- go to the library/use library database

Assess the quality of a journal article

Several rankings may help to assess the quality of research you consider

- Journals of general interest and by field in economics and management
 - For German speaking countries, consider the VWL / BWL Handelsblatt Ranking for economics and management
 - The German Association of Management Scholars provides an expert based ranking VHB JourQual 3.0, Teilranking Management im Gesundheitswesen
 - Web of Science Impact Factors
 - Scimago
- Health Economics, Health Services and Health Care Management Research: Health Economics Journals List
- Be aware that like in any other domain, there are predatory publishing practices.

Investigate how a journal article is connected to other works

- Citationgecko
- Connected papers
- scite_ – a tool to get a first impression whether a study is disputed or academic consensus

Organize your literature

- Zotero (free of charge)
- Mendeley (free of charge)
- EndNote (potentially free of charge via your university)
- Citavi (potentially free of charge via your university)
- BibTEX if you work with TEX
- Excel spread sheet

Example: Hellerstein (1998)

As an illustration of the research process of formulating a hypothesis, designing the study, running the study and collecting the data, analysing the data and, finally, reporting the study, we provide an example of the partly reproduction of a highly cited paper in the field of health economics . At the end of each chapter, we demonstrate insides into the study “The importance of the Physician in the Generic versus Trade-Name Prescription Decision” by Judith K. Hellerstein from the year 1998 (Hellerstein 1998) that we reproduced.

Relevance of the topic - escalating health expenditures

Worldwide, many health systems face constantly increasing health expenditures, especially for prescription drugs. In the United States, the total prescription drug expenditure in 2020 marked about 358.7 billion US-Dollars (Statista 2020). Prescribing generic instead of brand name drugs is an option in reducing the total health care expenditure. Generic drugs are bioequivalent in the active ingredients but are often less costly compared to their counterpart brand name alternatives. Discounts are as high as 80 percent (Pruckner and Schober 2018).

Development of a research question - identifying the role of physician practice in choosing generic drugs

Physicians are faced with a multitude of different medication options, including the choice between generic and trade-name products when prescribing the same active pharmaceutical ingredient. Physicians ideally act as agents for their patients to identify the best available treatment option. Choosing the best treatment entails cost of coordination and cognition. The prescription of generic drugs may serve as an example to what extent physicians customize treatments according to patient’s needs with regards to cost. From an economic point of view, once generic versions of is available a drug that has been previously subject to patent protection, a perfectly rational consumer would demand a generic drug instead of the trade-name version (Fischer and Stargardt 2016). Accordingly, Hellerstein investigates the following research question: Do *“physicians vary their prescription decisions on a patient-by-patient basis or whether they systematically prescribe the same version, trade-name or generic, to all patients”*.

The study focuses on two aspects:

1. identify the role the physician in prescribing a generic over a brand name drug

2. investigate role of a patient's insurance status in the choice between generic and brand name drugs by the physician

We will focus on the second aspect.

Hypothesis - insurance status drives generic drug use

To further investigate the stated research question and formulate a hypothesis, a motivation is needed why different groups by insurance status matter in the choice between generic and brand name prescription drugs to draw conclusions. Hellerstein (1998) considers a patient's insurance status as a matter of dividing the study population in groups for which the choice between generic and brand name drugs differs. The hypothesis to test the underlying hypothesis is *"Physicians are more likely to prescribe generics to patients who do not have insurance coverage for prescription pharmaceuticals"* (Hellerstein 1998).

Hellerstein (1998) hypothesises that there is a relationship between prescription of a generic drug (dependent variable) and insurance status (independent variable) of the patient is suggested.

Designing the Study

Basic steps in designing your study

We have arrived at the heart of your research study to specify how you will empirically test your hypothesis. We focus on study designs make use of secondary data with the aim to identify the magnitude of effects of causal relationships between a variable of interest X and a certain outcome Y , controlling for potential confounders Z . Note that there is extensive literature and guidance about designing and performing causal inference studies, some of which we guide you to below. We will highlight the most important elements and point to the relevant resources.

You have chosen your research questions and formulated a hypothesis. You also have collected and reviewed all the relevant literature on your topic. Now, you need to decide on your research design looks like. Are you planning to uncover associations, or do you want to examine a causal relationship between your variables of interest (Pearl 2009; Pearl, Glymour, and Jewell 2016)? We will focus on investigating causal relationships. To translate direct observations in data to investigate cause-and-effect relationships, you will need to rely on a workable model that describes the elements and relationships of concepts reflected in your hypothesis. For that reason, it is recommended that you first describe the causal structures of the elements that you are studying, including any other observed or unobserved structures that may disturb the cause-and-effect relationship under investigation.

The design of your study will depend on the causal model that suggests which effects are estimable, given that there are suitable data to empirically identify the effect. Consider that you describe cause-and-effect relationship first, assess how it is estimable (that means that you ask the question “Can you infer a causal effect from your data?”) and then investigate the effect by direct observations using secondary data. The alternative is to start out with a regression estimated effect and then investigate whether it has a causal interpretation.

Consider the subsequent steps as an iterative process, allowing that not all data you may need to estimate effects of your causal model will be readily

available. Some elements will remain unobserved. For others, you may need to find proxies.

Develop a causal model and identify estimable effects

You should specify a causal model that describes the relationships between the elements that you are studying. This causal model will help to identify empirical model by defining your main relationship of interest.

You need to specify your

- main outcomes variable of interest (X)
- main independent variable of interest (Y)
- other independent variables (confounders of the effect of X on Y), that means any additional Z 's
- unobserved variables (U)

You need to think about how these concepts can be measured and what the relationships between these concepts are.

For better visualization, you may plot a corresponding directed acyclic graph to describe the relationships between your variable of interest and any confounding variable. It is recommended to do this step before data collection. That way, you avoid collection of data that may not be needed. You also avoid forgetting variables that are necessary to for causal effect identification. And you may check for so-called bad controls that may bias your estimates.

Develop the empirical strategy to investigate causal effects

Potential outcomes framework notation

Write down your estimate, your estimand, notation

After you have specified your causal model, ask yourself how you can capture the causal relationship between X and Y ? In other words, can you apply an appropriate research design that can accommodate you in determining causality with the data at hand? Following the classification by Matthey et al. (2020), that means by performing, for example a

- Randomized Controlled Trial (RCT): this would be the ideal world case where we could randomize treatments to study effect

Secondary data will not allow performing a RCT as you will not be able to randomly assign subjects, such that you need to run a quasi-experiment to identify treatment effects using a

- Confounder Control Study
 - Regression adjustment
 - Matching Techniques
 - Simulation
 - Fixed effects regressions
- Instruments
 - Regression discontinuity design (RDD)
 - Difference-in-difference (DiD) analysis
 - Instrumental variable (IV) approach
- Any other method such as synthetic control groups, structural equation models, causal machine learning

Think about what type of treatment effect you are investigating in your causal analysis. The treatment effect is the average (across the population or across some subpopulation) of the change in outcome (Y) that results from a change in a covariate (the treatment X) (Lewbel 2019).

Common types of treatment effect parameters are

- average treatment effects (ATE)
- average treatment effects on the treated (ATT)
- marginal treatment effects (MTE)
- local average treatment effects (LATE)
- quantile treatment effects (QTE)

Investigate suitable source data

You need to find and collect suitable data to populate your causal model. This might take some time, and not all of the variables can be found in one dataset. You might frequently need to collect and combine a few different data sources. When searching for the data sources, you can turn to already existing datasets specifically designed for scientific use, for example surveys (NAMCS, CPS, NLSY, HILDA, RLMS HSE, KiGGS); panels (for example GSOEP, SHARE, HRS, ELSA); censuses (e.g. Microcensus). [INCLUDE LINKS]

Some data are available from statistical offices (Eurostat in the European Union, DESTATIS for Germany), private companies (health insurances), Social Media and App based data (Mappiness, Fitbit, Facebook, Twitter, WayBetter) or governmental and non-governmental institutions (EMA, FDA, BfARM). Sometimes you have to hand-collect and digitize the necessary data, for example from historical documents, data from governmental and non-governmental agencies, commercial providers and library search engines,

archives; download statistical tables from INKAR's or Unemployment Agency's websites. Therefore, you need to plan whether and how these can be accessed and how much time is needed for extraction. Consider automated tools like scraping.

You need to decide on what aggregation level you need to conduct your analysis to estimate the cause-and-effect relationships postulated, for example individual (patient, student), organization (hospital), regional (country, state).

Given that, you need to identify appropriate data sources that contain your variables of interest at the given aggregation level.

Resources box

Key terminology in causal inference based on Matthey et al. (2020)

- **Causal model:** A description, most often expressed as a system of equations or a diagram, of a researcher's assumptions about hypothesized known causal relationships among variables relevant to a particular research question.
- **Treatment, exposure, or independent variable:** The explanatory variable of interest in a study. Some also describe this as the "right-hand-side variable".
- **Outcome, dependent variable, or left-hand-side variable:** The causal effect of interest in a research study is the impact of an exposure(s) on an outcome(s).
- **Potential outcome:** The outcome that an individual (or other unit of analysis, such as family or neighborhood) would experience if his/her treatment takes any particular value. Each individual is conceptualized as having a potential outcome for each possible treatment value. Potential outcomes are sometimes referred to as counterfactual outcomes.
- **Exogenous versus endogenous variables:** These terms are common in economics, where a variable is described as exogenous if its values are not determined by other variables in the causal model. The variable is called endogenous if it is influenced by other variables in the causal model. If a third variable influences both the exposure and outcome, this implies the exposure is endogenous.

Openly-available textbooks and resources on econometrics with a causal inference focus

- Causal Inference: The Mixtape by Scott Cunningham
- What If by Jamie Robin and Miguel Hernan

- The Effect: An Introduction to Research Design and Causality by Nick Huntington-Klein
- How Do We Know if a Program Made a Difference? A Guide to Statistical Methods for Program Impact Evaluation by MEASURE Evaluation
- Open Source Economics Basics to microeconomic methods
- Introduction to Econometrics using R by Christoph Hanck
- Econometrics by Bruce E. Hansen
- Mastering Econometrics by Joshua Angrist
- Differences-in-Difference design, Health Care Policy Science Lab
- Statistical Tools for Causal Inference by the SKY Community

Directed Acyclic Graphs

- A more comprehensive overview of causal paths is provided by Nick Huntington Klein
- A tool to draw Directed Acyclic Graphs (DAGs) is DAGitty

Replication examples

- Replication of “Mostly Harmless Econometrics” in Stata, R, Python and Julia

Example: Hellerstein (1998)

With the identified research question and the hypothesis that physicians are more likely to prescribe generics to patients who do not have insurance coverage for prescription pharmaceuticals at hand.

Hellerstein sets up an observational, retrospective analysis of secondary data, including patient and physician characteristics. The goal of the analysis is to identify a causal effect of the patient’s insurance status (independent variable X) on the prescription of a generic over a trade-name drug (dependent variable, Y). As the prescription decision of the physician is prone to involve information imperfections and agency problems, the empirical model needs to control for physician specific effects among other possible influencing factors (confounders Z). Another factor, Hellerstein is concerned about is random variation in the choice between generic and trade-name drugs.

The confounder control study design chosen by Hellerstein (1998) uses a random effects probit specification at physician level to estimate the effect of insurance status on generic vs trade-name drug choice. This includes the variable of interest (estimate), a binary variable, which indicates whether a patient got a trade-name or a generic drug prescribed. The estimate takes the

value 1 if a generic drug or the value 0 if a trade-name drug was prescribed. Of interest is the causal effect of the patients different insurance types. Those include Medicare, Medicaid, HMO/prepaid and private insurance. Other independent variables included in the analysis for the purpose of reducing possible bias are the drug class as proxy for prices, patient characteristics (age, gender, race), physician specialist status, the region and the averages of patient characteristics per physician .

The choice of covariates to be included in a study depends on the study design, the relevant literature in the field and on possible restriction by the underlying data.

Often you will not find all variables that are potential confounders directly. Thus, your creativity is needed to find appropriate proxies, to use the data at hand to be able to account for any relevant possible biases. An example from

Hellerstein's (1998) study design is the need to include prices into the empirical analysis. Though prices are not provided in NAMCS, unobservable price differences are accounted for by including drug class dummy variables assuming that prices vary across different drug classes. This way, though we cannot capture variation in the prices of products directly, we account for possible bias of the different price level by drug classes.

Running the study and collecting the data

Basic steps when running the study and collecting the data

In this step it is about to specify which data to use and collect the data. As this study guide focuses on secondary data, you will need to identify relevant data sources, obtain the source data and process the data. As many data sources are not primarily developed for the particular purpose of your study, you often will need to modify the source data to reflect the measurements of your Y 's, X 's and Z 's you wish to examine in your causal model. It is also often that secondary data is not readily available for analysis such that you will need to collect and process these data (for example by writing program that extracts data from a website, search engine, or API, or by applying for permission to access the data).

We will distinguish between **source data set(s)** and the **analysis data set** that you will be using to run the study.

We point to the following distinction that will have implications on the resources needed to collect and process:

- Secondary source data that is readily available and has been documented earlier (for example panels)
- Secondary source data that you will need to process for the purpose of your study (for example by use of web-scraping, use of web-APIs, or extraction from data warehouses)

The following steps are typically performed when collecting the data:

Identify secondary data sources

Once you have designed the study, you can now search for appropriate data sources of your X 's, Y 's and Z 's. That means you investigate which data items of one or multiple secondary data source(s) you are using and how these items can be collected. It is part of this process that you document where, how and which data you gathered and processed. A codebook that includes a description of the content of the data source is such a piece of documentation. Specify exactly the variables that you will be extracting and any exclusions of the data source. If you are generating a dataset yourself from another source through scraping techniques, you equally need to document how the data are collected and processed, for example by sharing the corresponding script and a description of that data.

You need to assess the suitability of the potential secondary data sources. Types and use of secondary data sources has been described across different disciplines, for example Hair, Page, and Brunsveld (2019) for business data, Eriksson and Ibáñez (2016) for drug utilization studies, Fitchett and Heafner (2017) for social studies. It is important to assess the quality of the secondary data sources by criteria like the following (Hair, Page, and Brunsveld 2019):

- **measurement validity:** this may be difficult to assess, but you may want to look out for research that has used the data you aim to use as well
- **reliability:** is the source that you are using providing the data at sufficient quality (see more in Hair)
- **potential bias** arises when the data do not measure what you want to measure. Biases may arise from for example, changes in the sample from which the data is collected (for example if an institution changes the way data are reported, if the sampled population is changing)

Develop a data collection and analysis plan

- Use a data schema to describe how you combine different secondary data sets into your analysis data set.
- Even if there is only one source data set that you are using, you may need to think about how this data set needs to be processed to be fit for your final analysis purpose.
- Pay attention to how different data sources are linked, that means are there unique identifiers of the individual subjects that you are studying (for example identifiers for patients, physicians, firms, products). Are they equal across the different source data?
- Per each data set, describe which variables or items you plan to collect and how you aim to process the source data to fit the purpose of your study.

Obtain access to source data or generate the data

- Investigate how you can obtain access to the data including permissions, fees and any ethical considerations.
- Note that many data sources require registration and/or charge fees for obtaining access and processing requests. Special conditions for academics and students often apply. You may need to describe your research project and ask for permission. Allow these steps to take considerable time.

Process and prepare the source data

Once the source data has arrived, been extracted or simply you hit the download button, at this stage you need to develop a program (Stata do-file, R-code or code using other software) that documents the data extraction steps and that leads to an analysis data set. Make sure you save all raw source data in a safe place that in terms of data protection. Ensure, that you do not accidentally delete or overwrite these data.

When you develop the program, you will need to make decisions informed by the previous steps related to

- Combining, merging, appending datasets
- Specifying datasets and variables needed to estimate your regression model
 - Which variables are included/excluded?
 - Which variables need to be modified (recoded/calculated)? How?
 - Creating new variables
 - Labeling the variables
- Defining the study period
- Defining exclusion criteria (for example teens or adults only, general practitioners or physicians)
- Aggregating to the appropriate level (individual, family, county, state, country, patient, physician, or hospital, etc.)
- Specifying the final analysis data set: what outcomes (Y), variables of interest (X), confounders (other X 's), instruments (Z) are possible to use from the data?

Resources Box

How to organize your research project:

- Folder structure
- Tips, tricks and software for keeping research organized
- Organizing a research project

Coding practices: Best Practices for Data and Code Management in Projects

General principles on **How to code well**

- Make program files self-contained
- Use relative paths
- Identify inputs and outputs
- Automize
- Be consistent
- Comment and document
- Use spacing and indentation
- Do not substitute brevity for readability
- Beware of error causing codes (small caps, commas, semikolon)

Data cleaning

- Cleaning data in STATA

Collecting data from meta-data and web-scraping

- Introduction to web scraping: Resources

Example: Hellerstein (1998)

The empirical application of Hellerstein (1998) is based on three datasets as provided by the National Ambulatory Care Survey from the year 1989:

As the confidential data is key to identifying physicians and patients, for the purpose of reproducing Hellerstein's results, we resort to the publicly available data of 1991, that provides these identifiers in two publicly available datasets:

- NAMCS: contains demographic and medical information on the full sample of patients and is mainly used to introduce the analysis population
- NAMCSd: contains information on medications, that were mentioned, prescribed or given to a subsample of patients of the NAMCS data.

The publically available data of NAMCS can be downloaded at the Centers for Disease Control and Prevention (CDC). The data come in an .exe format. We use the unpacking software 7-zip to convert the files in a .txt format.

Usually, certain transformations need to be performed to be able to use the raw dataset. This can apply to the file or data format, the structure of the dataset or you might need to combine different datasets.

To process the raw data after we downloaded the raw data from CDC website, a challenge was to structure the raw data set to be able to use the relevant variables in Stata. When we imported the text file into Stata, all data was condensed to one string of characters in a single column. This means that we first need to indicate the different data items (variables) for the individual information (i.e. patient characteristics). Therefore, to be able to use the data, we need to manually split the string into the wanted format, so every information in a separate variable. Which digits in the initial string belong to one data item is defined in the documentation files (also often referred to as code book), provided by the CDC website. Almost every dataset comes with a codebook, containing useful information on the coding or handling of the data.

This can often help and provide the necessary information needed to successfully prepare your data for empirical analysis.

For the reproduction of Hellerstein (1998), we use Stata, creating a .do -file containing the code. It is of importance to develop all necessary steps of your research project to manipulate and analyze data using program files, containing every step you performed with your data. This ensures the reproducibility of your research and enables you to go back to earlier versions of your data preparation and analysis, if necessary.

After downloading the NAMCS and NAMCSd we performed the following steps to clean (by excluding non-relevant certain observations), transform and prepare the data for analysis. We concentrate on reproducing all tables and figures originally produced by Hellerstein (1998):

For NAMCS, used for the reproduction of table 1:

- Observations of patients, who were not charged for prescription pharmaceuticals, were dropped.
- Missing data must be relabeled according to the syntax of the statistic software and observations with missing data were not considered.
- The setup of the dummy variables for Medicare and Specialists underlie special conditions (details can be found in the do-file attached), whereas the remaining dummies were not created straight away.

For NAMCSd, used for all other figures and tables:

- Keeping only those observations of drugs that are the first mentioned drug in a patient visit.
- Keeping only those observations of drug mentions, that are part of the eight largest drug classes. The names of the included drug classes can be found in table 3 and the labels the data uses in the documentation files.

- Keeping only those observations of drugs, that were prescribed.
- Create the same dummies as in the preparation of the NAMCS data (foot-note table 2)
- Create an indicator on whether a drug is a multisource drug (definition given in ch.1 paragr. 1: “We categorized drugs with the same ingredients and define drugs in an ingredient-group as multisource if there is at least one generic and one tradename drug within a group.”).
- Dropping observations of medications that are not multisource drugs.
- Using the variable on whether a drug is a generic or trade-name drug to create an indicator on whether the drug is a generic (dependable variable).
- Creating physician averages of variables listed on p. 123 of Hellerstein (1998).

After completion of these steps, the analysis dataset included x variables and y observations. The dataset is now ready for statistical analysis (see at the end of the next chapter). Of importance is to ensure that the raw data, the data preparation .do -file and the analysis dataset are stored safely.

Analyzing the Data

Basic steps when analyzing the data

Think again about your research question and what you are trying to learn or discover. How can you use your data to answer this question?

The data to be analyzed should correspond to the core elements of the hypothesis to be investigated. You investigate whether the postulated cause-effect relationship exists, or quantitatively identify the strength of the effect. To perform a reproducible research project, you will need to create a set of programs that describes how the secondary data used was synthesized, process and analysed. To document all steps of your analysis, you should use programs like Stata or R.

Across all steps, try to use data visualization using tables and diagrams. These are an essential part of your work and not an accessory. Visualizations help to support the argumentation in the text and visualize complex facts in a simple form. In the following, we describe the three major tasks for data analysis

1. Generate the analysis data set
2. Describe your data, assess validity and plausibility
3. Generate estimates of your investigated effect using regression techniques

Generate the analysis data set

There are three major tasks that are typically needed to generate the analysis data set:

1. Clean your data.

- get/collect the data and transfer them in a format you use, for example to .dta from .xlsx (if necessary)
- Make sure that you link different data sets correctly using correct identifiers.

- Take time to look through the data, check them, and delete anything that looks suspicious.
- Select your sample of interest.
- Generate and leave only variables necessary for your analysis.
- Ensure that you make plausible and relevant exclusion decisions, for example regarding the time period studied, products, health conditions, age groups.

2. Structuring and aggregating of the analysis data

- Only include variables of your empirical model to be included to the analysis data set!
- Aggregate your data to the level of analysis. For example, if you aim to analyse physician behavior over time, there should be one observation for each time period and physician (panel data). If your data is purely cross-sectional, there should not be multiple time periods in your data. That means in a cross-section of physicians, one row represents one physician.

4. Store your data

- Your analysis data set is the most important piece for your analysis.
- Physically store the version of your analysis data set that you are using.
- Use version control if you are modifying your analysis data set.
- Ensure that the code to create your analysis data set is complete and can reproduce the analysis data set completely.

Describe your data, assess validity and plausibility

Start the data analysis by doing some descriptive analyses of your data and sample. Once you have selected the necessary variables, generated new variables for the analysis, and combined all the necessary datasets into one analysis dataset you can start your empirical investigation.

- Check the plausibility by looking at basic descriptives (N, mean, median, frequencies)
- Plot the distribution of your data using histograms, boxplots, bar charts
- Critically reflect any anomalies: Compare your data with the reference literature
- Are descriptives similar? If not, why so? Try to assess coding problems, different population, unbalanced samples.

Generate estimates of your investigated effect using regression techniques

- Once you have decided on the type of empirical analyses you would like to perform, run the regressions.
- Perform the regression analyses, that means by using the appropriate procedure to estimate regressions that reflect your empirical strategy.
- Create output tables that report your results which outsiders can understand, for example label your variables properly
- Concentrate on analyzing and interpreting the effects of your variable of interest (X)
 - interpret and show in tables only most important and relevant coefficients related to the research question. For example, you do not need to interpret and display coefficients for all of the included control variables, just the main variables of interest.
- Think about how to interpret your estimates.
- Challenge your approach. Investigate why your estimates could not be plausible?
- Again, compare with existing literature.
- What does your data say? After you run the regressions, look at the results:
 - Are your main coefficients of interest statistically significant, that means is the p-value smaller than 0.05 or any other pre-defined level of significance?
 - Consider also the economic significance of your results. Is the effect you are measuring large or small?
 - Do your results make sense or are they counterintuitive? This might give you a clue that you might have misspecified your regression or made a mistake in your analysis.

Resources box

Data Analysis

- Princeton University Library: Getting Started in Data Analysis using Stata and R
- Data Analysis for Business, Economics, and Policy

Organizing your workflow, programming and automation

- Gentzkow M, Shapiro J. Code and Data for the Social Sciences: A Practitioner's Guide [Internet]. Chicago Booth and NBER; 2014
- The Stata workflow Guide
- In Stata coding, Style is the Essential: A brief commentary on do-file style

Data Visualization

- The chapter Data Visualization Basics by Hans Sievertsen includes important resources for the fundamentals of data visualization
- Stata Cheat Sheet on Visualization provides an overview of the technical implementation
- Jones AM. Data visualization and health econometrics. Foundations and Trends in Econometrics. 2017

Create journal submission ready output tables

- Creating Publication-Quality Tables in Stata
- Stata commands to plot regression coefficients, make regression tables and visualize results by Ben Jann

Example: Hellerstein (1998)

Hellerstein's (1998) study aims to answer the research question if physicians vary their prescription decisions on a patient-by-patient basis or whether they systematically prescribe the same version of a drug, trade-name or generic, to all patients.

This means that secondary data (NAMCS and NAMCSd) that is publicly available via the websites of the CDC will be analysed on physician level in a cross-sectional format. Thus, the primary interest is to investigate if the physicians do or do not prescribe a generic drug. We are not investigating prescription behaviour of physicians over time using panel data. Over the course of your analysis, you want to explore your dataset, this includes describing and assessing the validity and plausibility. Thus before generating estimates of your investigated effect using regression techniques, you want to describe your data using descriptive statistics presented in tables and figures.

We first reproduce the descriptive statistics of Hellerstein (1998), namely table 1-3 and figure 1-2. Details of created variables can be found in the data preparation .do-file and the notes of the tables and figures in Hellerstein's (1998) paper.

Please keep in mind that we used the data of 1991. Thus, our numbers are not matching with the numbers provided in the original paper.

We explain the steps taken to reproduce table 1-3 and figure 1-2, and the reproduction of the empirical analyses. Details and the code can be found in the Appendix and at the end of the next chapter when reporting the study results.

Describe your data, assess validity and plausibility

Table 1:

Compute the mean and summary statistics for the age and the previously established dummies in the NAMCS data.

Table 2:

For the first two columns, repeat the procedure of table 1 using the NAMCSd data. For the third column, compute the mean of the generic indicator for all subpopulations defined by the dummies.

Table 3:

Using summary statistics command, count the number of observations for the full sample and for each of the eight defined drug classes. Further, compute the mean of the generic indicator in the full sample and each drug class subsample.

Figure 1:

For each physician, compute the mean of the generic indicator and drop duplicate observations per physician. This way, every physician has a unique observation with a respective share of generic prescriptions. To approximate the distribution shown in figure 1, we chose the following specifications based on visual approximation like:

- The decimal places at which the average generic share was rounded: .01
- The bandwidth: 0.02
- The kernel function: bilinear form

Figure 2:

To display Figure 2, keep physicians that prescribe the same multisource drug to at least six patients, independent of whether it is in its generic or trade-name form. Compute the mean of the generic share for each remaining physician and create a categorical variable on whether this mean is zero (only trade name drugs), one (only generic drugs), or something in between (both, that means a mix of generic and brand name versions). The frequencies of these three categories across physicians is then plotted in a bar plot.

Generate estimates of your investigated effect using regression techniques

Hellerstein applies a random effects probit regression model aggregated on physician level. The parameterization of the model is implemented as follows:

Dependent variable Y

- G : Generic compared to brand name drug use

Note that Hellerstein uses a somewhat different notation for the outcome variable which is denoted as G . Today, it is often common to denote this variable by Y .

Variable of interest or treatment X

- P : Insurance status by: Medicare, Medicaid, HMO/prepaid, private insurance, self-paid (this is the omitted category)

Note that Hellerstein uses a somewhat different notation for the variable of interest which is denoted as P . Today, it is often common to denote this variable by X or D .

Confounders Z of the effect of insurance status X and generic compared to brand name drug use Y .

- C : Drug class identifiers among 8 classes (Pain relief omitted, see footnote table 5)
- X : Patient characteristics: age, sex, race
- S : Physician specialist status: Specialist, general practitioner (omitted)
- R : Region as classified by: Midwest, South, West, Northeast (omitted)
- vector X , vector P : Averages \langle !-Of WHAT?- \rangle across physician

Note that Hellerstein uses a somewhat different notation for confounding variables. It is often common to denote the variable of interest using the index X , but not necessarily confounders.

Table 4 and 5:

Estimate the regression coefficients, t-statistics and marginal effects of the regression model. It is not specified whether the model for both tables contain all covariates specified in equation 10 or just the variables mentioned in the tables. We decided to use the full model for both tables. Marginal effects are the average marginal effects (AME) across individuals (footnote table 4).

Table 6:

Run a separate regression using only observations of a single drug class and report the coefficients of the payment dummies and their average marginal effects (AMEs).

Tables 7, 8 and 9 cannot be reproduced. The publicly available data of NACMS does not provide a variable to identify physicians and patients by region.

Reporting the Study

Basic steps to reporting the study

At this stage you need to report all your performed work in detail in a scientific research paper.

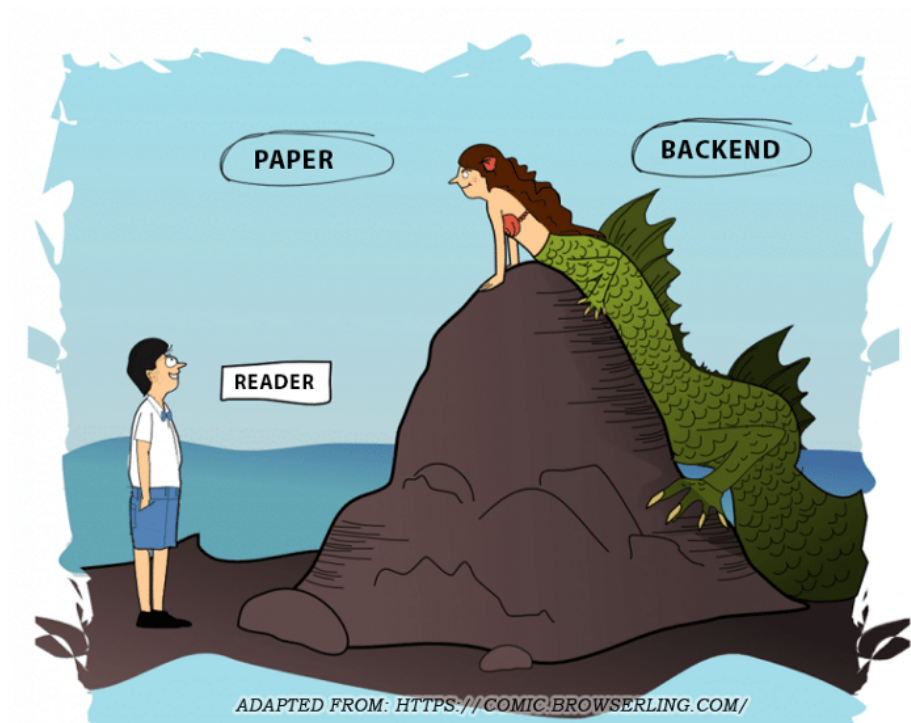


Figure 4: How to report your paper

Source: <https://towardsdatascience.com/how-to-keep-your-research-projects-organized-part-1-folder-structure-10bd56034d3a>

By reporting your study you provide readers with the information necessary for them to assess the contribution of your study and the soundness of methods used, allows them to reproduce the study if they wish, and decide whether they agree with the conclusions that you draw given your data, methods, and results. Thus, from the instructions from your paper a reader should understand and be able to reproduce every table and figure.

A research paper usually includes the following sections:

- Title
- Abstract
- Introduction
- Previous Literature Review
- Data and Variables
- Empirical Methods
- Results
- Discussion
- Summary and Conclusion
- References
- Appendices

We will briefly summarize each of these items.

The **Title** should be short (15 words or less) and reflect your research question. It needs to reflect the purpose (what is the question that you are addressing?), scope (i.e. if you cover a specific time period or population) of the study. Sometimes you might want to include the methods of your study in the title, if, e.g. you conduct a cohort study. All this information should help the reader to decide whether or not they are interested in reading it.

The **Abstract** is usually around 300 words and includes a very short summary of your research question, data and methods used, main findings, and conclusions.

The **Introduction** should be part of the paper that is written at the very end. It should motivate your research topic and give a more extensive overview of the context, data, methods and findings. You should provide context by referencing the existing literature regarding the things that we know and things that we do not know and then parlay the latter part into your research question. You need to interest your reader to a degree that they would want to continue reading the rest of your paper. The **Introduction** is said to be the most important part of the paper, as you have to grab the readers' attention.

The **Previous literature review** are standard in economic papers. They are either included in the **Introduction** or are placed in a separate section. It contains a critical review of the most relevant and influential (well-cited) studies on the topic. It should also contain a summary and a connection how

your study is related to what has been done before and how your study is going to contribute to it.

A significant part of your empirical project is conducting a thorough literature review. Once you choose your topic and pose a research question, one of the first things you usually do is look at the previous research and theories related to the research problem. By doing this, you will create a comprehensive overview of all the published knowledge on your topic available to date. This will provide a description, summary, background, context, relevance, and critical evaluation to the research idea you are working on. The main purpose of the review is to provide an overview of sources you are considering and to convince the reader of the need to conduct the research in question, and to show how it fits into the larger field of study. It will help you to establish whether the topic has been researched before, and show what problems others faced during their research.

In a literature review, you need to identify and summarize key scholarly publications from the fields that are pertinent to your research topic. It should involve a thorough search of the main key words in the relevant databases and journals.

The **Data** and variables section should contain a detailed information about the data source used in the paper as well as information about the variables used in the analysis.

Describe the data you are using in your study in great detail. Is it a secondary data? Did you collect it yourself, if yes, how? Describe your sample and variable selection.

The **Empirical method** section should in great detail describe the statistical method used in the paper and the assumptions that are necessary to be fulfilled in order to get an unbiased and consistent estimation of your parameter of interest. Also describe the limitations and how you addressed them.

Ideally, you should describe each step that readers should perform, if they would like to reproduce your results and conclusions.

The **Results** section should contain a set of tables and figures that show your empirical results. Make sure that tables are reader-friendly and are easy to comprehend. Place figures and tables as close as possible to the place in the text where you first referred to them.

Do not discuss or explain the results in this section, and do not provide any interpretation or speculation here. Keep it factual and simply describe in words what the tables display.

The **Discussion** section should contain the interpretation and discussion of the results, as well as potential mechanisms, and suggest a direction for future research.

You should start your discussion by reiterating your research question and based on your findings, state what you think an answer to your question is. Do not introduce any new data or results in this section. It should reflect only the results already presented in the paper. You need to interpret your results in the context of the literature that you identified and discussed in the Introduction and the Related Literature sections.

Are your findings consistent with what the other literature have found? Do your data fill the gap in the knowledge that you identified? Here you want to show how your work added to and extended the knowledge on the topic.

You can also discuss implications of your study, for example for public policy. You also should acknowledge and discuss the limitations of the study here and what implications these limitations might have for the results. The more accurate, open and detailed you are about the limitations, the more credibility your results will have. Offer any alternative explanations for your findings.

Discuss any negative effects in your findings and the impact they had on your conclusion. You can also talk about next steps and the implications of your findings for future research.

The **Summary and conclusion** section is usually short and contains the most important information on the topic, findings and concluding remarks. It should reflect main points mentioned in the Introduction.

Here you should underline why your research matters and state the answer to your research question. Also state here any recommendations that can be made based on your findings.

At the very end of your paper you need to provide a complete list of **References**. The bibliography includes a complete list of academic papers and books referenced and cited in your study.

The easiest way to keep track of your reference list is to write down the information, i.e. the title, author(s), journal/book, publisher, and publication date, about the original source each time you use it. Decide in advance on one citation and reference style and follow it throughout the paper.

The **Appendices** are a handy tool. In order to keep your main part of the paper focused, of reasonable length, and not overburdened, you can place all the still relevant, but secondary and less critical information, i.e. various robustness checks, descriptive statistics, etc., in appendices.

As a careful researcher you will conduct and document a number of various robustness checks of your main results. However, once you confirm and verified that your results are robust, it does not make sense to keep all of the robustness checks in the main body of the paper. Appendices allow you to save space. You can just summarize what you did and refer to the tables in the appendix. Some journals also provide possibilities of online appendix.

Source: Fresno State Graduate Writing Studio. Elements of a Research Paper.

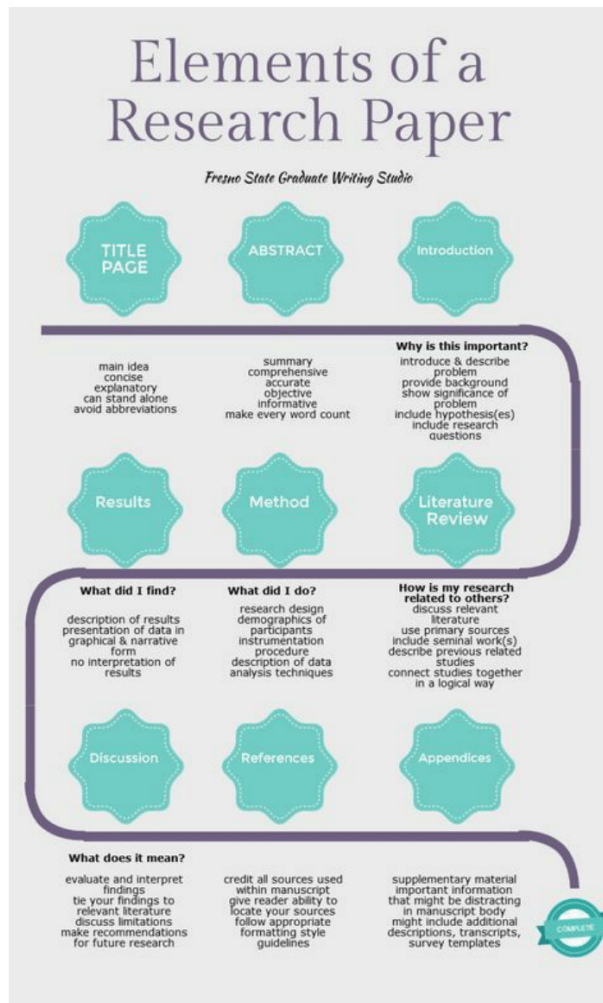


Figure 5: Paper Structure

The writing process

- producing a good empirical paper takes time. You should start working on it and drafting as soon as possible.
- allow yourself sufficient time for collecting and analyzing the data, writing and revising the paper. Writing a paper is a recursive process that will take many revisions.
- read the text yourself and let others read it. (“Is the text fun to read?”, “Would I want to read beyond the introduction myself?”, “Are the questions guiding the research comprehensible?”)
- keep it short. When editing ask yourself whether you can make the same point using fewer words?
- avoid repetition. Repeating things uses extra space and tests readers’ patience.
- use “I” in a single-author project, and “we” in the project with more than one author.

More principles on writing in economics are summarized in Top Ten Rules of Economical Writing and more comprehensively in (McCloskey and Ziliak 2019) and by (Hall 2013)

General guidance on the use of language in the research paper

- use present tense, that means “*in this paper I attempt to...*”
- generally stick to one tense
- use active voice
- use in-text citations, that means “*Hellerstein (1998) finds that...*” – even though it was a while ago; “*Table 1 shows...*”, not “*Table 1 will show...*”
- use simple and direct sentences. Keep them short.
- check your paper for repetition and cut whenever you detect it in your paper.
- try to be as clear as possible with a few words as possible.
- place minor or secondary details and digressions from the main paragraph into footnotes.

Tables

- each table should be self-explanatory
- use 2 to 4 digits after the comma when reporting the results, not all the numbers produced by the statistical program.
- show results with and without controls. For example, start with a column that includes only the main coefficient(s) of interest and then progressively

include various controls (for example patient characteristics, hospital FEs, etc.) in subsequent columns.

Figures

- use figures to show patterns in the data—they demonstrate it better than big tables with a lots of numbers,
- label the axes properly,
- provide self-explanatory captions.

Resources box

Collections of writing tips and short articles

- Varian, H. R. (2016). How to build an economic model in your spare time. *The American Economist*, 61(1), 81-90.
- Dudenhefer, P. (2009). A guide to Writing in Economics. EcoTeach Center and Department of Economics, Duke University.
- Nikolov, P. (2020). Writing tips for economics research papers
- Cochrane, J. H. (2005). Writing tips for Ph.D. students. Chicago, IL: University of Chicago.

Books

- Hall, G. M. (Ed.). (2013). How to write a paper (5th ed). Wiley-Blackwell.
- McCloskey, D. N., & Ziliak, S. T. (2019). *Economical Writing, Third Edition: Thirty-Five Rules for Clear and Persuasive Prose (Third Edition)*. University of Chicago Press.

Videos

- Greg Martin (2018). How to write a paper.

Example: Hellerstein (1998)

We show the results of the reproduction of tables 1-3 and figures 1-2 of the descriptive analysis, and the reproduction of the empirical analysis in table 4-6. Please keep in mind that our results are based on data from 1991. This in contrast to Hellerstein (1998) which is based on three datasets of the year

1989. As the confidential data is key to identifying physicians and patients in the data of 1989 we resorted to the 1991 data, that provide state identifiers in two publically available datasets. We present our results and discuss the differences compared to Hellerstein (1998) briefly. For a detailed discussion and interpretation, we refer to the original study.

Descriptive Analysis

The summary statistics shown in Table 1, representing the overall NAMCS patient sample of the year 1991 are similar to the summary statistics shown by Hellerstein (1998). When looking at the distribution of insurance coverage, we see differences especially at Blue Cross/Blue Shield (2 compared to 13 in 1989), other commercial insurers (38 compared to 24 in 1989) and HMO/prepaid plan (14 compared to 24 in 1989). Additionally, about 14 more specialist physicians are represented by the sample.

Table 1 - Summary Statistics for Overall NAMCS Patient Sample

Variable	Mean	Standard Deviation
Age	42.58	24.63
Female	0.59	0.49
Nonwhite	0.11	0.32
Hispanic	0.06	0.23
Self-pay	0.24	0.43
Medicare	0.12	0.33
Medicaid	0.09	0.29
Blue Cross/Blue Shield	0.02	0.15
Other commercial insurer	0.38	0.49
HMO/prepaid plan	0.24	0.43
Specialist	0.69	0.46
Northeast	0.23	0.42
Midwest	0.25	0.44
South	0.28	0.45
West	0.24	0.43
N	32407	

Notes: see Hellerstein (1998) Datasource: NAMSC91

Table 2 shows very similar statistics for the patients of the NAMCSd sample compared to the data from 1989. We see a 5 increase in the proportion of generic drug prescription in specialist compared to Hellerstein (1998).

Table 2 - Summary Statistics for Patients in NAMCS Drug Sample

Variables	Mean	Standard Deviation	Proportion Generic
Age	43.09	24.98	.
Female	0.589	0.492	0.27
Nonwhite	0.122	0.327	0.33
Hispanic	0.062	0.242	0.32
Self-pay	0.283	0.451	0.28
Medicare	0.128	0.335	0.22
Medicaid	0.101	0.301	0.32
Private	0.336	0.472	0.27
HMO/prepaid plan	0.142	0.349	0.35
Specialist	0.4	0.49	0.31
Northeast	0.208	0.406	0.28
Midwest	0.27	0.444	0.26
South	0.294	0.456	0.25
West	0.228	0.42	0.35
Full sample	.	.	0.28

Notes: see Hellerstein (1998), Datasource: NAMSCd91

Compared to 1989, we observe larger sample size in the 1991 data. Table 3 shows the absolute number of observations and the share of generics over all drugs as well as the 8 largest drug categories. While the absolute number of observations is higher, the share of generics overall and by drug class stays similar to Hellerstein (1998).

Table 3 - Frequency of Generic Prescription by Drug Class

Name of Class	Observations	% Generics
All drugs	14,008	29.20
Antimicrobials	4,236	43.77
Cardiovascular-renals	3,042	25.02
Central Nervous System	1,213	31.82
Hormones/Hormonal mechanisms	1,799	37.80
Skin/Mucous membrane	827	5.56
Ophthalmics	524	18.89
Relief of pain	1,711	11.63

Name of Class	Observations	% Generics
Respiratory tract	656	10.06

**Figure 1 - Distribution of Physician Generic Prescription Rates
(Source: NAMSCd91)**

The distribution of physicians over the different generic prescription rates (figure 1) is similar to Hellerstein (1998).

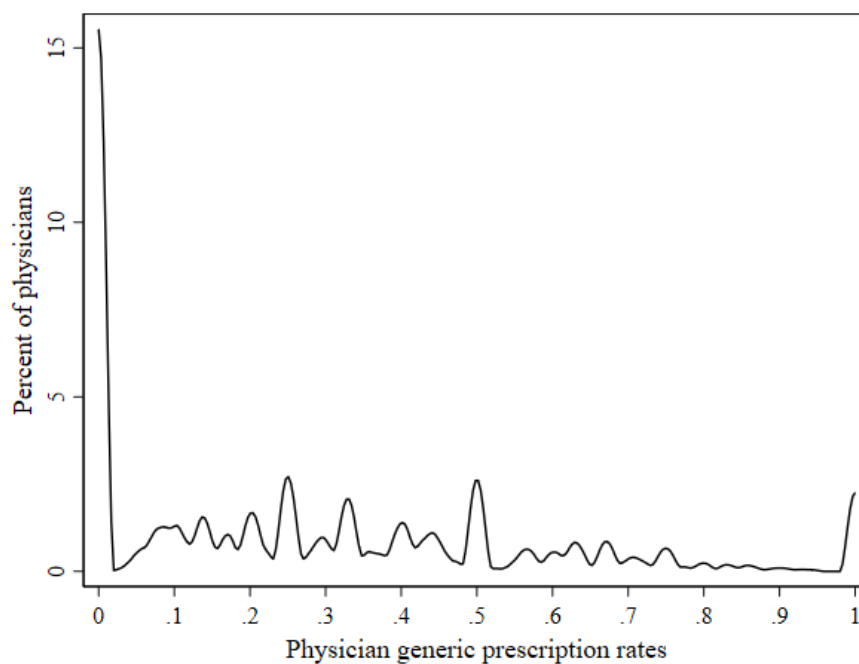


Figure 6: Generic prescription rates

Figure 2 - Physician decisions for physicians who prescribe a drug to at least six patients (Source: NAMSCd91)

When looking at figure 2, we see an decrease from about 90% to about 50% in “*only trade name*” prescriptions by physicians. This seems to be driven by a rather small increase of prescription of “*both versions*” and “*only generics*”.

Empirical Analysis

Please keep in mind, that, due to data restrictions, we were not able to control for certain covariates that Hellerstein (1998) included. An example of such are

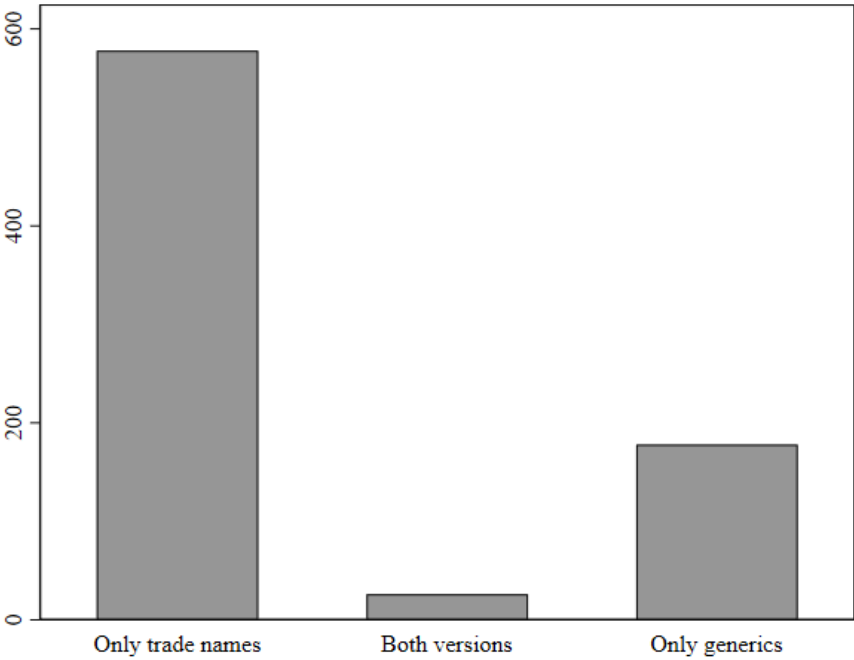


Figure 7: Physician decisions

legislation laws like mandatory or permissive substitution, and one- or two-line prescription. For this reason, we cannot present all tables produced by Hellerstein (1998). Given these restrictions, we cannot run the analysis with the same underlying model specifications. This restricts us from comparing the results of the years 1991 and 1989, however for the completion of this example we represent the results of table 4, 5 and 6 in the following:

Table 4 Estimated Coefficients on Demographic Variables, Geographic Variables, and Average Characteristics for Full Sample, excluding regional identifiers

Variable	Random-Effects Probit Coefficient	Random-Effects Probit t-statistic	% Change in Generic
Constant	0.71	0.94	-
Age	-0.01	-1.50	-0.002
Female	-0.16	-1.11	-0.0463
Hispanic	-0.54	-1.46	-0.154
Nonwhite	0.19	0.73	0.0534
Specialist	0.00	-	-
Mean age	0.01	0.54	0.0017
Percent female	-1.73*	-2.11	-0.494*
Percent black	-0.95	-1.52	-0.270
Percent Hispanic	0.45	0.29	0.127
Percent Medicaid	-0.01	-0.01	-0.0027
Percent Medicare	-2.53**	-3.01	-0.722**
Percent private	-0.99	-1.70	-0.282
Percent HMO/prepaid	-1.12*	-2.09	-0.321*

Notes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; see Hellerstein (1998);
Datasource: NAMSCd91

Table 5 - Estimated Coefficients for Drug-Class Dummy Variable for Full Sample

Drug Class	Random-Effects Probit Coefficient	Random-Effects Probit t-statistic
Antimicrobials	0.989***	4.25
Cardiovascular/renals	-0.231	-0.78

Drug Class	Random-Effects Probit Coefficient	Random-Effects Probit t-statistic
Central Nervous System	-1.118*	-2.03
Hormones/Hormonal mechanisms	0.280	0.87
Skin/Mucous membrane	-0.198	-0.44
Respiratory tract	-0.736	-1.29

Notes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; Omitted category is Ophthalmics; Notes: see Hellerstein (1998); Datasource: NAMSCd91

Table 6 - Tests for Moral Hazard for the Full Sample Equality of Individual Insurance Variables with Self-Payment Random-Effects Probit Results

Insurance Variable	Antimicrobials	Cardiovascular	Metabolic	Hormones	Skin/Mucous Membranes	Pain relief	Respiratory Tract
Medicaid							
Coefficient	0.178	-	-	0.497	0.189	-	0.634
t-statistic	1.49	0.0345	0.192	1.68	0.31	0.280	-1.017
% change	0.042	-0.15	-0.74	0.054	0.062	-0.74	-1.04
Medicare							
Coefficient	0.156	-	-	-	0.684	-	0.164
t-statistic	1.08	0.0397	0.388	1.07	1.15	0.515	-0.413
% change	0.047	-0.23	-1.43	0.103	0.017	-1.29	-0.50
Private							
Coefficient	-	-0.085	-	-	0.226	-	0.201
t-statistic	0.0012	0.170	0.053	0.27	0.59	0.394	-1.146
% change	-0.01	-0.54	-1.02	0.025	-0.036	-1.32	-1.42
HMO/Prepaid							
Coefficient	0.262	-0.156	0.0353	-	-0.395	-	0.362
t-statistic	2.29	-0.68	0.13	0.120	-0.72	0.238	-0.272
% change				0.42		-0.43	-0.29

Insurance Variable	Antimicrobials	Cardiovasculars	Metabolic	Hormones	Skin/Mucous Membranes	Pain relief	Respiratory Tract
% change	-0.000	-0.018	-0.034	-0.011	0.020	0.029	-0.096
					0.076		

Notes: see Hellerstein (1998); Datasource: NAMSCd91

Bezjak, Sonja, April Clyburne-Sherin, Philipp Conzett, Pedro Fernandes, Edit Görögh, Kerstin Helbig, Bianca Kramer, et al. 2018. *Open Science Training Handbook*. Zenodo. <https://doi.org/10.5281/ZENODO.1212496>.

Bollen, K., J. T. Cacioppo, R. M. Kaplan, Krosnick J. A., and J. L. Olds. 2015. "Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science." National Science Foundation. https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf.

"Developing Strong Research Questions Criteria and Examples." 2019. *Scribbr*. <https://www.scribbr.com/research-process/research-questions/>.

Eriksson, Irene, and Luisa Ibáñez. 2016. "Secondary Data Sources for Drug Utilization Research." In *Drug Utilization Research*, 39–48. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118949740.ch4>.

Fischer, Katharina Elisabeth, and Tom Stargardt. 2016. "The Diffusion of Generics After Patent Expiry in Germany." *The European Journal of Health Economics* 17 (8): 1027–40. <https://doi.org/10.1007/s10198-015-0744-3>.

Fitchett, Paul G., and Tina L. Heafner. 2017. "Quantitative Research and Large-Scale Secondary Analysis in Social Studies." In *The Wiley Handbook of Social Studies Research*, 68–94. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118768747.ch4>.

Hair, Joe F., Michael Page, and Niek Brunsveld. 2019. "The Nature and Sources of Secondary Business Data." In *Essentials of Business Research Methods*, 4th ed. Routledge.

Hall, George M., ed. 2013. *How to Write a Paper*. 5th ed. Chichester, West Sussex: Wiley-Blackwell.

Hellerstein, Judith K. 1998. "The Importance of the Physician in the Generic Versus Trade-Name Prescription Decision." *The RAND Journal of Economics* 29 (1): 108–36. <https://doi.org/10.2307/2555818>.

"How to Write a Strong Hypothesis Steps and Examples." 2019. *Scribbr*. <https://www.scribbr.com/research-process/hypotheses/>.

Huschka, Denis. 2013. "Why Should We Share Our Data, How Can It Be Organized, and What Are the Challenges Ahead?" *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2272028>.

- Lewbel, Arthur. 2019. "The Identification Zoo: Meanings of Identification in Econometrics." *Journal of Economic Literature* 57 (4): 835–903. <https://doi.org/10.1257/jel.20181361>.
- Matthay, Ellicott C., Erin Hagan, Laura M. Gottlieb, May Lynn Tan, David Vlahov, Nancy E. Adler, and M. Maria Glymour. 2020. "Alternative Causal Inference Methods in Population Health Research: Evaluating Tradeoffs and Triangulating Evidence." *SSM - Population Health* 10 (April): 100526. <https://doi.org/10.1016/j.ssmph.2019.100526>.
- McCloskey, Deirdre N., and Stephen T. Ziliak. 2019. *Economical Writing, Third Edition: Thirty-Five Rules for Clear and Persuasive Prose*. Third Edition. Chicago ; London: University of Chicago Press.
- Meyer, Klaus E., Arjen van Witteloostuijn, and Sjoerd Beugelsdijk. 2017. "What's in a P? Reassessing Best Practices for Conducting and Reporting Hypothesis-Testing Research." *Journal of International Business Studies* 48 (5): 535–51. <https://doi.org/10.1057/s41267-017-0078-8>.
- Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. "A Manifesto for Reproducible Science." *Nature Human Behaviour* 1 (1): 1–9. <https://doi.org/10.1038/s41562-016-0021>.
- Orozco, Valérie, Christophe Bontemps, Elise Maigné, Virginie Pignet, Annie Hofstetter, Anne Lacroix, Fabrice Levert, and Jean-Marc Rousselle. 2020. "How to Make a Pie: Reproducible Research for Empirical Economics and Econometrics." *Journal of Economic Surveys* 34 (5): 1134–69. <https://doi.org/https://doi.org/10.1111/joes.12389>.
- Pearl, Judea. 2009. *Causality*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511803161>.
- Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. Chichester, West Sussex: Wiley.
- Pischke, Steve. 2012. "How to Get Started on Research in Economics?" http://econ.lse.ac.uk/staff/spischke/phds/get_started.pdf.
- Pruckner, Gerald J., and Thomas Schober. 2018. "Hospitals and the Generic Versus Brand-Name Prescription Decision in the Outpatient Sector: The Generic Versus Brand-Name Prescription Decision." *Health Economics* 27 (8): 1264–83. <https://doi.org/10.1002/hec.3774>.
- Schilbach, Frank. 2019. "5 Steps Toward a Paper." MIT 14.192 guest lecture. <https://www.dropbox.com/s/q7wjaidl5w91srt/Guest%20lecture%20FS.pdf?dl=0>.
- Statista. 2020. "Prescription Drug Spending in U.S. 1960-2020." <https://www.statista.com/statistics/184914/pre-scription-drug-expenditures-in-the-us-since-1960/>.

- Trochim, William. n.d. "Problem Formulation." Accessed May 13, 2021.
<https://conjointly.com/kb/problem-formulation/>.
- Varian, Hal R. 2016. "How to Build an Economic Model in Your Spare Time."
The American Economist 61 (1): 81–90.
<https://doi.org/10.1177/0569434515627089>.
- Vilhuber, Lars. 2020. "Reproducibility and Replicability in Economics."
Harvard Data Science Review 2 (4).
<https://doi.org/10.1162/99608f92.4f6b9e67>.