INSTITUTE FOR DEFENSE ANALYSES

Leveraging Machine Learning in Defense Personnel Analyses

Julie Lockwood
Alan Gelder
Matthew Goldberg
Jennifer Brooks
George Prugh

May 2020
Approved for public release;
distribution is unlimited.
IDA Paper P-13174
Log: H 20-000152

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

About This Publication
This work was conducted by the Institute for Defense Analyses under contract HQ0034-14-D-0001, project BE-6-4311, "Assisting implementation of and analysis within DOD's Enterprise Data to Decisions Information Environment," for the Office of the Under Secretary of Defense for Personnel and Readiness (OUSD (P&R)). The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The authors thank John "Jay" Dennis, Michael Guggisberg, Nancy Huff, Karen Gordon, and Peter Levine for their careful review and helpful comments, and Jordan Marcusse for his contributions at an early stage of this paper.

For More Information:
Dr. Julie A. Lockwood, Project Leader
jlockwood@ida.org, 703-578-2858
ADM John C. Harvey, Jr., USN (Ret), Director, SFRD
jharvey@ida.org, 703-575-4530

INSTITUTE FOR DEFENSE ANALYSES

IDA Paper P-13174

Leveraging Machine Learning in Defense Personnel Analyses

Julie Lockwood
Alan Gelder
Matthew Goldberg
Jennifer Brooks
George Prugh

This page is intentionally blank.

Executive Summary

Recent developments in computing and data-driven algorithms have brought “advanced analytics” to an ever-growing array of topic areas. Leaders in government and industry are constantly bombarded with opportunities to apply machine learning and other “big data” techniques that offer to improve their products, institutional management, or competitive edge. Those seeking useful analyses will benefit from a better understanding of what the various techniques can provide and what they cannot.

Advanced analytics are not magic. To obtain meaningful insights, one must carefully determine the foundational analytic construct, select the right tools for the job, apply those tools to appropriate data, and remain aware of any limits on how results can be interpreted or applied. Fancy math does not guarantee scientific validity. This paper describes the primary applications of advanced analytics to policy questions, organized by the type of question asked into *what?*, *how? (*or *why?)*, and *what if?* analytic categories. Institute for Defense Analyses researchers illustrate this framework using defense personnel research questions, but the same structure can be applied to nearly any topic area. Although we focus on recently developed machine learning tools, traditional statistical tools can also address questions in each class.

“What?” Questions: Descriptive Analysis and Forecasting

“What” questions seek to describe—not explain—past or present conditions or to forecast future conditions. Analyses in this class use existing patterns to generate predictions, groupings, classifications, and tabulations and to identify trends and correlations. Machine learning techniques excel in this area, often producing predictions at much higher levels of accuracy and fidelity than other methodologies. This increased precision flows from the algorithms’ systematic, agnostic approach to identifying complex relationships and patterns within data. The natural caveat for all forecasts is that these methods implicitly assume that the patterns and relationships observed in the past are representative of the future.

Some examples of “what” questions facing the military services include:

- What are the current recruiting and retention patterns for each military service? For specified subgroups? What is expected to occur over the coming years? Where are shortfalls projected in the near future?

- Of a given applicant pool, which individuals will likely succeed in basic training, special forces training, pilot training, or another program? What individual characteristics observed before beginning training predict with success?
- What are the early warning signs for depression, post-traumatic stress syndrome, or other mental health challenges, and which service members have those signs?
- What units are likely to maintain strong cohesion and morale? Where are morale or disciplinary problems likely to arise? What characteristics of leaders, units, teams, missions, and individuals are correlated with maintaining strong cohesion and morale or with morale or disciplinary problems?

Questions of this type center on describing what exists, has existed, or will probably exist—not why. As descriptions and forecasts, answers to these questions alert leaders to trends and conditions. Detailed forecasts of retention patterns, for example, can highlight potential personnel shortages, and leaders might adjust the volume and timing of their recruiting and training pipelines to fill these anticipated gaps. However, the forecasts cannot explain *why* retention rates are expected to increase or decrease, or *if* a given policy would change things. Forecasts and other pattern-based analyses can instead provide hypotheses for further examination.

"How?" (or "Why?") Questions: Retrospective Causal Analysis

After forecasting what is likely to occur, attention naturally shifts to *why* the forecast is thus or *how* one can influence the expected outcome. Backward-looking analyses that learn about cause-and-effect from events or conditions in the past can provide some answers. Determining cause and effect is now the central focus of the foundational analytic construct. Causal analysis tries to identify how an outcome would change if one or more of the contributing conditions changed by measuring the effect of different policies and events on outcomes to help guide policy makers and measure return on investment (ROI).

Some examples of "how" (or "why") questions facing the military services include:

- How did the introduction of the Blended Retirement System (BRS) impact retention of mid-career officers in each service, and by how much?
- How has the Navy's Assignment Incentive Pay (AIP) program impacted the retention of individuals serving in billets that are difficult to fill?
- How has the Army's Career Satisfaction Program (CSP) impacted officer recruiting?
- How did an increase in reenlistment bonuses change retention rates in a given career field?
- Why did one area meet its recruiting goals, while another area did not?

- Why do some individuals perform better than others in certain roles?

To answer a how (or why) question, machine learning algorithms must be implemented within a causal model (i.e., within a scientific framework that shows how different environments, policies, processes, or conditions affect workforce outcomes). The causal model contextualizes the environment or policy and the resulting outcomes and provides a basis for hypothesizing about causal relationships.

Causal models rely on experiments, which can be either intentional or natural. A classic intentional experiment is the randomized controlled trial, where different individuals are assigned to different treatments at random. However, intentional experiments are often impractical or infeasible. Fortunately, many mathematical tools developed in the econometrics and statistics disciplines allow analysts to evaluate outcomes under various conditions as natural experiments and thus uncover causality. Natural experiments can exist if, for example, a certain policy is more likely to apply to some people than to others for reasons that are essentially random (after controlling for other systematic differences). In many cases, machine learning algorithms can be used in experimental frameworks to provide modeling flexibility to improve performance.

"What If?" Questions: Prospective Analysis

What if no experiments or quasi-experiments exist? Relevant real-life observational data are often scarce or non-existent, and experiments may be impossible or unethical. Forward-looking "what if" hypothetical questions ask what will happen in the future if important changes occur (e.g., if a major policy that has never been tried is introduced). The foundational analytic construct for "what if" questions requires a blend of causal analysis and structured forecasting. Pure forecasting is not reliable because the underlying conditions affecting peoples' decisions will have changed, and pure retrospective casual analysis is impossible because the situation is novel and no data exist.

Some examples of "what if" questions facing the military services include:

- What will reenlistment rates be among various subsets of individuals if Selective Retention Bonuses (SRBs) are set at a given level?
- Would extending the Navy's AIP program to selected positions or to other military services improve retention in high-demand areas?
- If Congress enacts a new retirement benefit policy, what is the expected impact on personnel retention at various grades and levels of personnel quality?
- If Navy sea tours lengthen or sea-shore rotations rebalance to extend sea duty, what would be the impact on fleet performance, morale, health, and retention?
- How would retention outcomes likely shift if the typical four-year enlistment contract is replaced with a variety of one- to six-year contracts?

- Would replacing or restructuring a particular career field impact immediate or future combat support effectiveness?
- If a large civilian labor market shift occurred, what impacts would it have on short- and long-run recruiting and retention and in which career fields?

Estimating the impact of a proposed policy requires explicit and valid assumptions about how individuals, groups, or organizations make decisions in different environments. However, traditional tools cannot handle the complexities of the human decision-making process. For example, existing models of military members' continuation decisions focus almost exclusively on compensation—ignoring career histories, familial considerations, deployment propensities, assignment locations, and other factors—because computational complexities have prohibited a more realistic representation of individuals' decision frameworks and preferences. Machine learning can dramatically expand researchers' ability to capture the intricate dimensions of military service and human decision making. These advances offer significant promise for expanding the scope of "what if" questions that can be examined in defense personnel analysis. Investment is needed to develop necessary tools for answering *what if* defense personnel policy questions. Just as the Department of Defense (DOD) invested in developing the basic science needed to apply dynamic programming to defense personnel policy questions in the 1970s and 1980s—resulting in a method that has been used for a generation—so it should again lead a new generation of analytic progress by investing to mature the application of machine learning advances to structural policy analyses and modeling.

Recommendations to achieve high-quality analyses for the DOD

High-quality, actionable research requires carefully framing the research question, choosing the right analytic tools for the job, applying those tools to rich and reliable data, and accurately interpreting the results. Creating the conditions for this success requires sustained investments. Personnel and other data assets must be available and maintained at a high level of quality. Mechanisms must be in place to securely bridge and combine data assets across different organizations and missions within DOD. Algorithms used in personnel analyses must be open to critical peer review to enhance their performance and to guard against legal, moral, and ethical abuses. The following list summarizes the sustained investments that we believe are needed for long-term success in applying machine learning and other advanced techniques to defense personnel analyses:

- Reduce barriers to establishing data access and sharing across organizations within DOD and between DOD and other entities.
- Harmonize data to enable the linking of information across organizations.
- Establish secure analytic computing environments where analysts in DOD and those who support DOD can access data, work, and collaborate effectively.

- Create and implement a pipeline for transforming frequently used “raw data” items into “research ready” cleaned, standardized, documented data objects in a reproducible manner.
- Ensure that data are preserved over long periods of time to allow for studies of the entire relevant process.
- Collect and preserve information on all choices and benefits offered to individuals—not just information on what option(s) an individual ultimately chooses
- Improve understanding of what kinds of questions can be answered by different algorithms and models.
- Invest in developing and sustaining reusable toolkits of algorithms and data, especially those that can be applied across military services and agencies. This investment should include the following:
  - Develop reusable algorithms for modeling military members’ continuation decisions that more realistically capture the multi-faceted decision process.
  - Adapt a suite of algorithms from the juncture of machine learning and econometrics for enhancing analyses of previously enacted policies.
  - Build forecasting tools for predicting the likelihood that individuals will successfully complete training programs or career milestones.
  - Mature tools and methods for deploying forecast results to appropriate DOD leaders at meaningful levels of granularity, together with a feedback loop for documenting actions taken in response to the forecasts.
- Encourage an open-source, collaborative culture among creators and users of analytic tools throughout the DOD enterprise.
- Adopt best practices for auditing the results of processes using operationalized machine learning or artificial intelligence tools. Actively investigate the legal, moral, and ethical implications of using the results of such algorithms.

Applied thoughtfully, advanced analytics such as machine learning, econometrics, and structural modeling can contribute significantly to DOD’s effectiveness. Many new tools fall in the interdisciplinary junction of computer science, statistics, and econometrics. This rapidly evolving area offers significant promise for generating insights to effectively cultivate and efficiently manage a high-quality workforce. To make the most of these advances and avoid mistaken or misleading research interpretations, leaders should ensure that the analytic methods applied are appropriate to the questions that they ask and are applied to the right data. The framework outlined here can help.

This page is intentionally blank.

Contents

This page is intentionally blank.

1. Introduction

Recruiting, developing, and retaining a high-quality
military and civilian workforce is essential for warfighting success.[1]

– 2018 National Defense Strategy

Recent developments in computational processing and data-driven algorithms have coalesced to bring “advanced analytics” to data large and small in an ever-growing array of topic areas. Leaders in government and industry are constantly bombarded with opportunities to apply machine learning and other advanced analytic techniques that offer to improve their products, institutional management, or competitive edge. Leaders seeking to obtain useful analyses will benefit from a better understanding of what the various techniques can provide and what they cannot.

Advanced analytics are not magic. To obtain meaningful insights, one must select the right tools for the job, apply those tools to the appropriate data, and remain cognizant of the underlying scientific method and constraints. This paper assists leaders by describing the three main applications of advanced analytics to policy questions, organized by the type of question the leader needs answered. We illustrate this framework using examples from defense personnel management research, but the structure can be applied to nearly any subject. We specifically describe how machine learning algorithms can be used to generate new or better insights in the following three question families:

- **“What?” questions.** These questions seek to describe the past or present or to predict future conditions. Example: What are the demographics of an organization’s workforce at present, and how will the demographic composition likely change over the next five years, assuming that current conditions persist?
- **“How?” (or “why?”) questions.** These backward-looking questions seek to explain or attribute cause and effect to events or conditions in the past. Example: How (or why) did a policy impact the success or failure of a workforce objective and by how much?

[1] Department of Defense, *Summary of the 2018 National Defense Strategy of the United States of America: Sharpening the American Military’s Competitive Edge* (Washington, DC: Office of the Secretary of Defense, 2018), 7, https://dod.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf.

- **“What if?” questions.** These forward-looking questions seek to project what will happen in the future if some aspect of the current environment changes. Example: What if a different retirement benefit policy is enacted? What kind of an impact will it likely have on a given workforce objective?

By providing answers to these questions, advanced analytics such as machine learning, econometrics, and structural modeling can contribute significantly to an organization’s effectiveness. In the context of defense personnel management, answering these questions can enhance readiness. To build a more lethal force, the National Defense Strategy emphasizes the need to deliberately cultivate a talented workforce. This requirement compels the Department of Defense (DOD) to “use information, not simply manage it.”[2] DOD tracks detailed information on millions of uniformed and civilian employees, including each individual’s assignments, deployments, skills, education, family structure, and other features. Managing these data is an ongoing challenge. Using them to inform and improve the management and performance of the force is difficult but not impossible.

Machine learning algorithms have expanded the breadth of tools available for obtaining insights on recruiting, training completion, retention, and other workforce management measures across each of these question families. Compared to traditional regression techniques, machine learning captures the interactions within vast troves of data in a much more systematic—rather than ad hoc—fashion. The researcher, for instance, does not need to hypothesize up front all the different subpopulations within the military that are likely to reenlist at different rates due to a deployment experience. Systematically capturing complex and often unexpected interactions within the personnel data allows for a higher degree of fidelity in understanding the state of the force. It moves the analysis from the coarse to the specific (e.g., from the cohort level to the individual level). Uncovering these interactions also enables an assessment of their relative importance to the question at hand because superfluous data can be separated from essential data. The ability to identify relevant data combinations can improve the tractability of some high-dimensional, computationally intensive problems that would otherwise choke on the so-called “curse of dimensionality.” However, machine learning is not a panacea. Like other analytic tools, these algorithms are appropriate for addressing some—not all—questions and are at various developmental stages ranging from embryonic to somewhat mature.

Table 1 contains examples of *what*, *how* (or *why*), and *what if* questions pertaining to the military workforce. As a hypothetical example of all three, *what* provides retention forecasts for service members within a particular occupation over the next five years, *how* (or *why*) provides an assessment of how a policy enacted eight years ago on select military bases to make daycare options more accessible to service members has impacted retention

[2] Ibid., 8.

Table 1. Examples of What, How (or Why), and What If Questions

What?	
Descriptive analysis or forecasting	What are the recruiting and retention patterns for each military service? What is expected to occur over the next three years?
	In what career fields are shortfalls projected in the near future?
	What are the projected trends for recruiting and retaining women in the force?
	Which applicants will likely succeed in basic training, special forces training, pilot training, or another training program?
	Which service members are likely experiencing early warning signs for depression, post-traumatic stress syndrome, or other mental health challenges?
	What units are likely to have strong cohesion and morale?
	Which career fields have similar recruiting profiles?
	Which first-term service members are likely to renew their contracts?
How? (or *Why?*)	
Retrospective causal analysis	How has the Navy's Assignment Incentive Pay (AIP) program impacted the retention of individuals serving in billets that are difficult to fill?
	How has the Army's Career Satisfaction Program (CSP) impacted officer recruiting?
	How did an increase in reenlistment bonuses affect retention rates in a given career field?
What If?	
Prospective analysis	How would a change in the retirement vesting timeline impact the number and quality of individuals available as field-grade officers?
	What would happen to personnel retention if certain career assignments were lengthened?
	Would replacing or restructuring a particular career field impact immediate or future combat support effectiveness?

outcomes, and *what if* examines how retention outcomes could shift if a future National Defense Authorization Act replaced the typical four-year enlistment contract with a variety of one- to six-year enlistment contracts. Retention is the unifying thread throughout these three cases. The objective, however, shifts from obtaining a retention forecast (similar to a weather forecast), to assessing the specific impact on retention of a previously enacted policy (perhaps to gauge whether the policy impacts merit the cost), to exploring the potential retention effects of a policy that has not been enacted and for which, consequently, no data exist on the direct effects.

The first class, *what,* is purely descriptive, using patterns resident with the data to generate predictions, groupings, classifications, and tabulations and to identify trends and correlations. For example, what does the current workforce look like? What will it look like in three years, and where are shortfalls likely to occur? Where are morale or disciplinary problems likely to arise? As forecasting tools, machine learning techniques provide DOD leaders with a way to obtain predictions, often at much higher levels of fidelity than previous methodologies. This increased precision flows from the algorithms' systematic, agnostic approach to identifying complex relationships and patterns within data. The natural caveat is that these methods implicitly assume that the patterns and relationships observed in the past are representative of the future. Chapter 2 discusses how DOD can better harness its data with machine learning techniques to improve its visibility into the current and projected state of the force.

Since the focus of *what* questions centers on the forecast itself (not why a forecast may be trending up or down or what would happen if a policy changed), there is no need to specify the potential channels of cause and effect. With *how* (or *why*) or *what if* questions, determining cause and effect is the central focus. Causal analysis tries to identify how an outcome (e.g., retention) would change if one of the conditions contributing to the outcome (e.g., the accessibility of daycare options) changed and by how much. The degree to which the outcome changes in response to a contributing condition is the *causal effect* of that condition on the outcome.

Understanding causation is critical in measuring the military's return on investment (ROI) from offering service members various benefits and bonuses. For instance, if the Selective Retention Bonus (SRB) for service members with a given skill set increases by $5,000, how much does that increase affect the likelihood that those service members will reenlist? Would they have already chosen to reenlist, or does that extra $5,000 tip the scales in favor of reenlistment—and by how much? If that group of service members was already inclined to reenlist, would the extra $5,000 make a decisive difference for some other group of service members? Causal analyses measure the effect of different policies and events on outcomes.

Machine learning algorithms excel at pattern recognition within data, but, without a causal scientific framework, they cannot explain *how* (or *why*) those patterns arose within the data in the first place. A forecasting algorithm may identify that service members in a particular population group have a high propensity to exit the military after four years of service, but it does not provide insights into the root cause for that pattern. To answer a *how* (or *why*) question, machine learning algorithms need to be augmented with a *causal model* (i.e., a scientific framework that shows how different environments, policies, processes, or conditions affect workforce outcomes). The causal model contextualizes the environment or policy and the data produced under it. Due to cultural, systematic, or other factors, was a policy likely to impact certain people more than others? How broadly was a policy implemented, and how was it applied? Context informs the interpretation of the data and provides a basis for hypothesizing about causal relationships.

A classic setting for understanding causality is a randomized controlled trial, where different individuals are assigned to different treatments at random. Often, experiments are impractical or infeasible to carry out. Thankfully, many mathematical tools developed in the econometrics and statistics disciplines allow analysts to evaluate outcomes under different policies or conditions as natural or quasi-experiments and thus uncover causality. Throughout the text, we provide examples for how these classic tools relate to military personnel analyses and, for further context, we review a small selection of previous studies on military career decisions in Appendix A.

Increasingly, researchers are combining traditional causal modeling frameworks with machine learning algorithms to exploit large or complex data and provide richer causal answers. We describe some of these advances—and how they relate to defense personnel analysis—in Chapter 3.

The third class of questions (*what if*) seeks to identify the effect of hypothetical policies, events, or practices that leaders might consider implementing. The challenge for these types of questions is that relevant real-life observational data are often scarce or nonexistent. Anticipating the impact of a proposed policy requires explicit and valid assumptions about how individuals, groups, or organizations make decisions in different environments. For example, such a model may assume that a military member will renew her service contract if her expected compensation value of renewing outweighs her other career options. Modeling this assumption requires using measurable factors to approximate the expected value of renewing the contract and the expected value of the alternative options available from not renewing the contract. Monetary compensation, while important and objectively measurable, is not the only factor that matters. It cannot capture the unique lifestyle, commitments, and experiences that characterize military service. However, existing models of military members' continuation decisions focus almost exclusively on compensation—ignoring career histories, familial considerations, deployment propensities, assignment locations, and other factors—because computational complexities have prohibited a more realistic representation of individuals' decision frameworks and preferences. The integration of machine learning techniques into structural models may loosen these constraints, as we discuss in Chapter 4.

Our taxonomy of *what*, *how* (or *why*), and *what if* is similar to Pearl and Mackenzie's description of moving from *seeing* to *doing* to *imagining*. Seeing is descriptive and associative, doing implies investigating the impact of an intervention to achieve some outcome, and imagining delves into the world of thought experiments and counterfactuals.[3] Heckman has also developed a taxonomy of causality, which splits the *how* (or *why*) portion between

[3] Pearl and Mackenzie use this terminology in their "ladder of causation." See Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (New York, NY: Basic Books, 2018).

identifying causation that is *internally* and *externally valid*.[4] Studies of internal validity entail identifying the impact that a policy has had on a group that has experienced that policy. However, that impact may be unique to a particular population, culture, or point in time. Whether that impact transcends from one context to another is a question of external validity.

In the chapters that follow, we review the machine learning tools that are available for addressing *what*, *how* (or *why*), and *what if* questions. Many of these tools are new, and they lie at the interdisciplinary junction of computer science, statistics, and econometrics. This rapidly evolving literature offers significant promise for generating the insights needed to effectively cultivate the high-quality military and civilian workforce that our nation needs. However, advancing meaningful analyses using these tools requires sustained investments. Personnel and other data assets must be available and maintained at a high level of quality. Mechanisms must be in place to securely bridge and combine data assets across different organizations and missions within DOD. Algorithms used in personnel analyses must be open to critical peer review to enhance their performance and to guard against legal, moral, and ethical abuses. In Chapter 5, we summarize conditions needed to create long-term success in advancing the use of machine learning in defense personnel analyses.

[4] James Heckman's taxonomy of causality moves from the *how* (or *why*) question of internal validity to the *how* (or *why*) question of external validity and, from there, to questions of *what if*. He does not include the observational *what* questions. See James J. Heckman, "The Scientific Model of Causality," *Sociological Methodology* 35, no. 1 (August 2005): 1–97, https://doi.org/10.1111/j.0081-1750.2006.00164.x.

2. “What?” Questions: High-Fidelity Prediction

Looking in the rear-view mirror is a bad way to drive.
Machine learning, on the other hand, is applicable to
datasets where the past is a good predictor of the future.[5]

– François Chollet, 2018

DOD, as the largest employer in the United States, maintains expansive data on its nearly three million current employees (2.15 million service members and 730,000 civilians), and historic data on millions of former service member and civilian employees.[6] These personnel data are often administrative, such as regular finance records, special bonuses, assignment locations, deployments, familial information for insurance coverage, and occupational skills. Other personnel data track performance on assigned duties, fitness tests, training requirements, commendations, citations, and disciplinary events. Data on health and wellness, intellectual aptitude, and various psychological attributes further augment the characteristics of service members. These data enable the identification of patterns in and contributors to many career and performance outcomes.

These DOD personnel data are dispersed across numerous organizations. Selected administrative data for the active duty, reserve, and civilian workforces are centrally collected and maintained by the Defense Manpower Data Center (DMDC), but the most extensive training, aptitude, health, performance, and readiness data are maintained and controlled by the military services. The Defense Health Agency (DHA) also maintains health records. Surveys on external perceptions and the internal state of the force are conducted within the Office of People Analytics (OPA) (e.g., by the Joint Advertising, Marketing, Research, and Studies (JAMRS) group). In the National Guard, the National Guard Bureau tracks some data elements centrally, but many more are only recorded by service elements within the individual states and territories, with meaningful differences in record-keeping. Selective training programs, such as those for special operations or pilots, often keep detailed performance logs on candidates as they move through a taxing battery of

[5] François Chollet, *Deep Learning with Python* (Shelter Island, NY: Manning Publications, 2018), 224. François Chollet is the author of the Keras, an open-source neural-network library written in Python and one of the foremost libraries for constructing neural networks and performing deep learning.

[6] Current employment numbers are from “Our Story,” U.S. Department of Defense, accessed December 9, 2019, https://www.defense.gov/our-story/.

tests but do not share this information outside of the program office. Although each organization's data can be used in isolation to answer some questions about the current or predicted state of the force, joining data across organizations can provide a more comprehensive picture.[7]

Restricting data access to isolated pools significantly impedes evidence-driven performance improvement. For instance, predictive forecasts of readiness that use administrative data from the DMDC but omit relevant health data from the DHA or training data from a military Service lack key information that could enhance forecast quality. Many questions, such as measuring the interdependent flow of individuals with certain training specialties between active duty, reserve, and civilian roles, can be fully examined only by viewing the DOD as an integrated system. When interdependencies exist, focusing on a single isolated aspect misses the broader system-wide effects. System-wide data are required to analyze system-wide outcomes.

Given DOD's wealth of data, machine learning techniques are required to provide actionable insights that fully leverage the Department's knowledge stores. In particular, these techniques can improve the accuracy and fidelity of forecasts that are needed for effective force management and program planning. An example is the selection of individuals for uniformed military service. DOD spends vast sums training, accommodating, disciplining, reassigning, removing, and replacing individuals who are ill-suited to military service. High-quality forecasts of early career milestone success could be developed using information available before formal enlistment. By identifying correlates to success or failure, this approach can help the services avoid accessing those with a low chance of success or identify areas in which innovation might improve success rates (e.g., by providing different pre-boot camp physical training). Budgeting for specialized career training programs—often done far in advance—provides another example. Such budgets are based on the number of individuals who are expected to enter that career field to meet future staffing needs. Improved retention forecasts for those currently in the career field can inform training budgets.

Forecasting is a primary example in the *what* question class. The remainder of this chapter describes the technical aspects of the major machine learning forecasting tools and provides additional examples.

A. Overview of the Supervised Learning Process

In forecasting, we seek to predict a future outcome or event as accurately as possible. The method should be *replicable*. Similar inputs should lead to similar outputs. The method should also be *externally valid*. It should perform well when tested on data inputs not used

[7] In a similar vein, the predictive value of DOD data is enhanced when these data are combined with information from other government agencies, statistical organizations, industry, and other sources.

in its development. Since forecasting's emphasis is on generating accurate predictions—not on identifying causal relationships—forecasts may be based on correlative patterns that have little underlying meaning or interpretation.

Machine learning algorithms can be grouped by the amount of information that is provided about the outcome that we desire to predict.[8] *Unsupervised* algorithms look for natural clusters among the data inputs when an outcome is not specified. They group like observations together. *Semi-supervised* algorithms require some—but not much—information on an outcome. When the outcome is known for a small percentage of observations, missing outcome data can be inferred using clusters formed on the data inputs. *Supervised* algorithms require complete information on an outcome. For instance, DOD has data that record the skills and career experiences of service members who have either chosen to reenlist or who have separated from the military. The inputs (the skills and experiences) and the outcome (reenlistment or separation) are observable for service members who have reached the decision point. Supervised machine learning algorithms attempt to identify the patterns that link the inputs to the outcome. If the link between the inputs and the outcome is stable, then the algorithm can use the learned patterns to predict outcomes using data from new inputs. In our example, an algorithm can use the skills and experiences of service members to predict whether they choose to reenlist before they reach the decision point. Because DOD personnel data frequently contain the outcome of interest, we will focus on supervised machine learning algorithms.

The usefulness of a supervised machine learning algorithm for forecasting hinges on its external validity. Suppose that the goal is to forecast the likelihood that service members reenlist. A valid forecasting algorithm must be able to predict reenlistment decisions that occur after the date of the data used to form the prediction, using data that were not involved in building the predictive algorithm itself.[9]

External validity can fail for at least two major reasons.[10] First, the algorithm may identify patterns between the inputs and outcome that exist in the data used to develop the algorithm but that do not exist generally. Capturing these coincidental patterns is known as *overfitting*. Second, a fundamental shift in the world may alter the connection between the

[8] Outcomes might include whether a service member reenlists or successfully completes a training milestone or whether the service member's unit meets a given level of mission readiness.

[9] We call these data "out-of-sample" data because they are excluded from the estimation process. The process of dividing the data into in-sample and out-of-sample components is called *sample splitting*.

[10] Numerous other potential failures have been documented. One taxonomy divides machine learning failures into *unintentional failures*, such as those described here, and *intentional failures*, whereby the data or algorithm is the target of a malicious attack. Attacks may attempt to do such things as fabricate results, alter the underlying data, or reverse engineer the algorithm or underlying data. See "Failure Modes in Machine Learning," Microsoft, November 2019, https://docs.microsoft.com/en-us/security/failure-modes-in-machine-learning.

inputs and outcome in the past and those in the future. In our retention example, the occurrence of a war, economic downturn, or new technology might break the historic relationship between observed characteristics and reenlistment decisions. Although accounting for such fundamental shifts is difficult, well-established procedures are available to reduce the risk of overfitting.

In supervised learning, the data are split into *training*, *validation*, and *test set*s. The underlying equation describing the relationship is estimated using the training set. This "learning" step matches patterns in inputs, such as an individual's skills and experiences, to outputs, such as a reenlistment decision.[11] The estimated model is then applied to predict the validation set outcomes, and these predicted outcomes are compared to the actual outcomes. That is, the estimated model is evaluated against out-of-sample data that were not used to build the model, which allows the researcher to see how well the estimated model performs in forecasting outcomes in new data. Since the algorithm is built from the training data, we expect it to be biased toward patterns that occur in the training data, hence the need to evaluate the model's performance on data not used to build it.

Machine learning algorithms also contain *hyperparameters*—aspects of the underlying mathematical model chosen by the researcher but not derived from data. In *hyperparameter tuning*, the training data are used to evaluate different hyperparameter choices. A group of models that are produced using different hyperparameters are compared by validating their performance on the test set. The model with the best performance on the test set is then selected as the model to use in the forecasting exercise.[12]

Estimating the model on the training data results in a bias toward patterns that exist within the training data. Similarly, model selection by use of the validation data can introduce bias toward patterns in the validation data. To overcome these problems, researchers reserve a third portion of the data, referred to as the test set. The final model selected through the process described previously is then evaluated against the test set.

B. Example: The Retention Prediction Model (RPM)

For many questions concerning retention policy, yes/no reenlistment outcome analysis does not provide enough information to assess retention across the full career spectrum. Retention forecasting over longer time horizons is better reflected by the time-to-event framework of survival analysis. Just as epidemiology studies examine how long patients are likely to survive after the onset of a disease and engineering studies estimate how long

[11] The machine learning literature uses the term "feature" to refer to an input. Other disciplines use terms such as variable, covariate, or data field to mean the same thing. We use these terms interchangeably.

[12] This model selection process can be extended by constructing multiple training/validation data splits. Again, the best model, or an average or hybrid of the models, can be selected for use. This model selection exercise underlies the procedure referred to as *cross-validation*.

a machine is likely to operate before failing, military leaders seek to forecast how long service members will remain in the force. Survival analysis allows information to be harnessed from observations that have not yet experienced the end state. Even if a patient has not died, a machine has not failed, or a service member has not left the military, the fact that they have persisted as long as they have carries information that can be used to predict how long similar observations will last.

The Institute for Defense Analyses (IDA) Finite-Interval Forecasting Engine (FIFE) incorporates time-to-event survival analysis into a machine learning context. This suite of algorithms can flexibly adapt to a broad variety of time-to-event settings. IDA's RPM applies the FIFE to predict the likelihood that a service member—at any point in his or her career—will remain in service through any future point. These individual forecasts can be aggregated based on any service member characteristics to get a group-level forecast for any portion of the force (e.g., by military service and occupation). High-quality predictions of this type enable DOD leaders to better anticipate where staffing shortfalls are likely to occur and to target retention efforts at high-value performers who are most likely to leave service.

The RPM considerably outperforms the traditional method that the military services use for predicting retention and the more sophisticated survival analysis implemented without machine learning. Roughly 12 percent of service members leave active duty service each year. A perfect forecasting algorithm would exactly identify 100% of exiting service members. The 12 percent of individuals with the highest forecasted probability of leaving should exactly match the list of those who actually leave. The military services typically rely on historical continuation rates at the occupation by year-of-service level to predict retention patterns. Using that method, the 12 percent of individuals with the highest forecasted probability of leaving only include 33.4 percent of actual leavers. A proportional hazards models increases that number to 35.7 percent. By contrast, IDA's RPM, which is built upon basic administrative personnel records, currently identifies 61.4 percent of those who will actually leave in its top 12 percent most likely to leave. Ongoing work to incorporate additional data and model refinements will likely increase the RPM's accuracy even further.

The RPM implements machine learning algorithms from two major families: tree-based models and neural networks. Subsections B.1 and B.2 give an overview of each family.

1. Tree-Based Models

Tree models split data into successively finer groups that share common outcomes. For instance, in predicting who will succeed in special operations training, one split (a *tree*) may divide individuals by cognitive test scores above and below a threshold. Each of these splits could then be split again (into *branches*), perhaps dividing the high test score subset on whether they come from a cold or warm climate and dividing the lower test score subset by their orienteering skills test performance. These cuts are determined automatically by the model. The groupings produced by a sequence of splits depends on the order in which the splits are made. The objective of tree models is to identify the splits and split orderings that result in the best model predictions. Algorithms for these *simple decision trees* were developed by Leo Breiman et al.[13]

A prominent shortcoming of simple decision trees is that the splits may only reflect coincidental characteristics of the data, compromising the tree's ability to make externally valid predictions. Breiman's *random forests* method[14] overcomes this drawback by creating enough variety in the trees to separate genuine predictive features from spurious ones. Each tree is constructed from a random set of features that can be used to split the data, and the data for each tree are restricted to a random subset of the total available data. This process produces variation across the trees. The predictions of each individual tree are then combined to create a single overall prediction.

An alternative approach sequentially builds trees on top of each other, with each successive tree using the portion of the data that the previous tree did not explain. This unexplained portion of the data, known as the *residual*, is the difference between the actual values and the predicted values. In Jerome Friedman's concept of *gradient boosted trees*,[15] each successive tree attempts to minimize the residuals according to some specified *loss function*.[16] Unlike other tree models, building trees on top of each other allows the full set of observations to be considered repeatedly—permitting rich, overlapping interactions. Gradient boosted trees methods have recently dominated competitions (e.g., those on kaggle.com/competitions) for generating predictions with tabular data. The ability to capture these interactions matters when modeling complex human decision making because much of the military personnel data are in tabular format, with a row for each individual and columns for the characteristics of the individuals.

[13] Leo Breiman et al., *Classification and Regression Trees* (Boca Raton, FL: Chapman & Hall/CRC, 1984).

[14] Leo Breiman, "Random Forests," *Machine Learning* 45, no. 1 (2001): 5–32, https://link.springer.com/article/10.1023/A:1010933404324.

[15] Jerome H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics* 29, no. 5 (October 2001): 1189–1232, https://www.jstor.org/stable/2699986.

[16] Mean squared error and absolute error are example loss functions. To speed the search for a model set that minimizes the loss function, the algorithm moves in the direction of the steepest downward slope. The steepest downward slope is called the *negative gradient* and is one of the namesakes for gradient boosted trees. To facilitate moving in the direction of the negative gradient of the loss function, the residuals, which serve as the input data for the next tree, are constructed in such a way that they approximate the negative gradient. Boosting, the other namesake, refers here to the sequential process of building predictive trees one after the other, with each using the residuals from the previous tree.

2. Neural Networks

Different types of data benefit from different analytic methods. In tabular data, the order of the data is not informative. Moving a row of administrative data that describes one service member to before or after a row that describes another service member does not change the information about either person. However, reordering words, pixels, movie frames, or sound bites can alter or destroy meaning for data media such as text, images, video, or audio recordings. When the order of the data matters, *neural networks* are the leading tool—although they can also perform well in other contexts.

Many potential applications of this modeling type to personnel questions exist. For example, neural networks could be used to identify depressed individuals through analysis of vocal recordings. Existing psychological questionnaires often fail to identify early warning signs of depression in men because these inventories rely on self-reporting. Recent studies suggest that more accurate diagnoses are possible by evaluating voice recordings.[17] Symptoms of depression that are detected using neural network analysis of voice recordings could provide better diagnoses among military service members and veterans and enable early and proactive help to prevent tragic outcomes such as suicide. In the accessions process, such analyses could improve the psychological health evaluation of potential recruits.

Neural networks are built using collections of decision points called *neurons*. A neuron takes one or more inputs, each with an associated weight and an overall bias, and transforms the inputs through an activation function, which, analogous to neurons in the brain, represents whether a neuron contributes to the outcome examined. Many activations combine to produce an overall prediction. Neurons are arranged in *layers* and accomplish different types of data processing. Generally, an initial layer of neurons processes the original data inputs. The outputs of this first layer are passed to the next layer of neurons, whose output can be fed into further layers of neurons and so on until the outputs of the final layer are consolidated into an overall prediction. Each neuron in each layer adds a degree of flexibility in assessing the inputs. Neural network design requires creativity, with the researcher choosing the number of layers, the number of neurons per layer, the activation functions, and the placement of connections between layers.

[17] See, for example, Ellen. W. Mc Ginnis et al., "Giving Voice to Vulnerable Children: Machine Learning Analysis of Speech Detects Anxiety and Depression in Early Childhood," *IEEE Journal of Biomedical and Health Informatics* 23, no. 6 (November 2019): 2294–2301, https://ieeexplore.ieee.org/document/8700173; Rachelle Horwitz-Martin et al., "A Vocal Modulation Model with Application to Predicting Depression Severity," in *Proceedings of the 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)* (Stoughton, WI: The Printing House, 2016), 247–253, https://ieeexplore.ieee.org/document/7516268; Nicholas Cummins, Julien Epps, and Eliathamby Ambikairajah, "Spectro-Temporal Analysis of Speech Affected by Depression and Psychomotor Retardation," in *2013 IEEE International Conference on Acoustics, Speech, and Signal Processing: Proceedings* (IEEE, 2013), 7542–7546, https://ieeexplore.ieee.org/document/6639129.

The possible connections between layers are vast and varied. Layers can be *fully connected* (or *dense*), with the outputs of every neuron in the preceding layer fed as inputs into every neuron in the following layer, or the output from any neuron in one layer can be connected to any number or selection of neurons in the subsequent layer. During training, the machine will process each set of features (representing a single observation) and make a prediction. Depending on the accuracy of its prediction, the algorithm will adjust the weights within neurons to improve performance on subsequent observations. Weights are updated once the machine has processed a *batch* or subset of the full set of training observations. Neural networks typically make several complete passes through the full set of training observations. Each full pass is referred to as an *epoch*.

Neural networks have several variants. *Feedforward neural networks* (FFNs) strictly move outputs from one layer to the next. In FFNs, all neurons in a layer process the data simultaneously, independent of other neurons in the same layer. The algorithm learns interactions between features as multiple neurons from one layer feed information into the same neuron (or neurons) in the next layer. *Recurrent neural networks* (RNNs) predict a sequence of outcomes using previous predictions in a sequence of inputs. For example, an RNN may be used to score the free-text performance evaluations of candidates being considered for promotion by identifying sequences of words that predict whether past candidates had been promoted.[18] A variant of RNNs, called *Long Short-Term Memory* (LSTM), replaces the traditional neuron with a small collection of decision points called a memory cell. The memory from one cell is passed to the next memory cell in the same layer, which enables longer short-term memory than a basic RNN. The RNN and LSTM architectures have the property that inputs occurring earlier in a sequence have less influence over inputs that occur many time steps later. *Attention mechanisms*, commonly used in natural language processing tasks, address this problem by enabling the algorithm to decide at any time step which other time steps to pay attention to.

C. Forecast Uncertainty

Predictions generated through machine learning are often characterized by a single point, such as the probability of continuing in military service over a given time horizon or the magnitude of a projected staffing shortfall. The uncertainty of the estimate provides information on risk. If a group of service members is predicted to remain in service for six years on average, it matters whether that average is mostly made up of service members leaving after five to seven years or after two to ten years. To calculate uncertainty bounds around an estimate, one must understand the sources of uncertainty. Basic sources of

[18] To help maintain consistency in the evaluation process, promotion boards or other processes that entail evaluating numerous candidates could be augmented by machine learning algorithms. Such algorithms could be used as an independent reviewer so that an arbitrator could take a closer look when discrepancies arose between the algorithm’s evaluation and the human evaluation.

uncertainty include how well the model represents reality, how well the particular data sample represents the broader population, and the uncertainty related to random chance (that can be modeled by an error term).[19]

Vast literatures in statistics and econometrics provide many methods for quantifying uncertainty. Traditional methods rely on the underlying properties of a model to calculate uncertainty bounds.[20] To the extent that these underlying properties approximate the modeled data process, the uncertainty bounds are mathematically valid. *Bootstrapping* provides a method for quantifying uncertainty without making assumptions about its true statistical distribution. The bootstrapping method takes random subsamples of the original data (with replacement), produces an estimate for each subsample, and then constructs uncertainty bounds based on the distribution of estimates. In the machine learning context, a researcher could take several random subsamples of the data, train a model with each subsample, produce estimates with each model, and then construct uncertainty bounds based on the distribution of estimates. However, this process is often too time intensive. For machine learning models that take hours or days to train, it may be computationally infeasible to train thousands of such models. The *bag of little bootstraps* technique introduced by Kleiner et al.[21] provides a scalable alternative that helps to alleviate this obstacle.[22]

While some traditional methods (e.g., the bootstrap) can be adapted to machine learning, uncertainty measures for machine learning forecasts are currently far less developed than for traditional statistical settings. Athey and Imbens observe that the requirements to produce for valid measures of statistical uncertainty “often come at the expense of predictive performance”[23]—the primary metric in machine learning. Some recent

[19] Several taxonomies exist for different types of uncertainty. One taxonomy breaks uncertainty down into epistemic uncertainty (relates to things that model does not capture but could be learned and incorporated into the model) and aleatory uncertainty (pertains to random chance).

[20] For instance, simple linear regression models assume that errors are independent and identically distributed.

[21] Ariel Kleiner et al., “A Scalable Bootstrap for Massive Data,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 76, no. 4 (September 2014): 795–816, https://www.jstor.org/stable/24774569.

[22] Instead of performing the bootstrap algorithm on massive data, the bag-of-little-bootstraps approach takes considerably smaller samples of the data, and the bootstrap algorithm is performed on each sample. Reducing the volume of the input data speeds the run time for the machine learning algorithm. If small-enough samples are taken, the time to complete a single run decreases by orders of magnitude. These smaller runs can form the basis of the bootstrap result.

[23] Susan Athey and Guido W. Imbens, “Machine Learning Methods Economists Should Know About,” *Annual Review of Economics* 11 (August 2019): 694, https://doi.org/10.1146/annurev-economics-080217-053433.

research explores uncertainty in specific machine learning contexts.[24] Academic developments should be monitored to capture uncertainty concepts that are meaningful for defense personnel applications.

[24] See, for instance, Yarin Gal and Zoubin Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning,” in *Proceedings of the 33rd International Conference on Machine Learning* – Volume 48, ed. Maria Florina Balcan and Kilian Q. Weinberger (JMLR.org, 2016), 1050–1059, http://proceedings.mlr.press/v48/gal16.pdf; Tim Pearce et al., “High-Quality Prediction Intervals for Deep Learning: A Distribution-Free, Ensembled Approach,” in *Proceedings of the 35th International Conference on Machine Learning* – Volume 80, ed. Jennifer Dy and Andreas Krause (Red Hook, NY: Curran Associates, Inc., 2018), 4075–4084, http://proceedings.mlr.press/v80/pearce18a/pearce18a.pdf; Tim Pearce et al., “Uncertainty in Neural Networks: Approximately Bayesian Ensembling,” arXiv:1810.05546v5, (2019). An earlier example is Tom Heskes, “Practical Confidence and Prediction Intervals,” in *Proceedings of the 9th International Conference on Neural Information Processing Systems*, ed. Michael I. Jordan, Yann LeCun, and Sara A. Solla (Cambridge, MA: The MIT Press, December 1996), 176–182, http://papers.nips.cc/paper/1306-practical-confidence-and-prediction-intervals.pdf.

3. "How?" (or "Why?") Questions: Retrospective Policy Analysis

Causality is a very intuitive notion that is
difficult to make precise without lapsing into tautology.[25]

– James Heckman, 2006

Data do not understand causes and effects; humans do.[26]

– Judea Pearl, 2018

Recruiting, developing, and retaining an effective workforce requires policies and practices that enable that workforce to thrive. In evaluating existing policies and shaping new ones, leaders benefit from understanding the differential impact that each policy has on desired workforce outcomes. The direct impact of a policy is often unclear—perhaps leading to positive results for one segment of the workforce and negative results in another. The magnitude of any impact on results can also differ across areas. Identifying the direction and magnitude of cause-and-effect relationships requires a guiding scientific model, a well-delineated course for testing that model, and valid measures for inputs and outputs.

The classic method for establishing causality is the *randomized control trial* (or *experiment*) where individuals or objects are randomly assigned to *test* and *control* status. To examine the effects of a new policy, for example, the policy could first be implemented on a randomly selected set of military bases (the test group), while all remaining military bases continue to use the existing policy (the control group). Randomization of assignments is meant to construct groups whose only meaningful difference is the treatment that they receive, allowing the researcher to attribute differences in outcome to the tested treatment. The researcher can reasonably infer that the treatment *caused* the outcomes to be different, which is usually not the case if individuals can choose whether to be in the test or control group (as is common in many social science contexts). Random assignment, rather than

[25] Heckman, "The Scientific Model of Causality." James Heckman received the Nobel Prize in Economics in 2000. He has worked extensively to develop and employ econometric methods for identifying causality.

[26] Pearl and Mackenzie, *The Book of Why*. Judea Pearl was awarded the 2011 Turing Award, the top honor in computer science "for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning" (see "Judea Pearl," ACM A.M. Turing Award, http://amturing.acm.org/award_winners/pearl_2658896.cfm).

self-selection, prevents individuals from gravitating to one group or another due to some underlying factor (e.g., their skills, family background, peer group, socio-economic status, and so forth). Without random assignment, inferring a causal relationship between treatment and outcome is methodologically challenging.

Absent intentional experimental conditions, causal inference is often based on events that arguably mirror experimental conditions for at least some portions of the population of interest. Such situations are known as a *quasi-experiments* or *natural experiments*. For example, a newly enacted state law affecting the availability of child care in that state—but not in the neighboring states—may provide a natural experiment for examining the effect of the availability of child care on the retention of service members with young children. Such a study would also need to account for all other relevant conditions that affect retention patterns for service members in each state.

Arguing that experimental conditions are met requires assumptions about the data-generating process (which should be tested). These assumptions must include a scientific theory that specifies the channels of cause and effect. The theoretical framework can be simple or complex but must answer the counterfactual question: "How would the outcome be different if the policy were not in place?" The outcome could be a measure of recruiting, training, retention, or readiness. The treatment can be binary (a policy is in place or it is not), a matter of degree, or encompass distinct options (e.g., differing bonus levels).

A causal claim is easier to validate within the population that experiences the experiment or quasi-experiment since individuals' responses to the policy change can be observed in the data. The applicability of a causal claim to populations and conditions essentially identical to those in the environment studied is known as *internal validity*. The applicability of an experiment or quasi-experiment's results to outside populations or conditions is known as *external validity*. Extrapolation from some individuals' experiences to anticipate how others will react is not always valid, since cultural, technological, or time-specific factors relevant for the experimental setting may exist that do not extend to other populations. For instance, a causal pattern that exists among enlisted individuals may not apply to officers, or a causal pattern that existed among reservists before 9/11 may not apply to reservists who entered service during the subsequent military engagements in Iraq and Afghanistan. On the other hand, if the same causal pattern is observed among enlisted and officer or both pre- and post-9/11, it strengthens the argument that the causal pattern applies broadly (across time, space, culture, or other relevant dimensions).

Many tools are available for establishing causal relationships and understanding whether the results hold generally. We briefly highlight a few of these methodologies from the econometrics and statistics literature, describe how machine learning techniques can augment them, and explore why one might want to use machine learning in this context.

A. Overview of the Traditional Causal Toolkit

Under classic experimental conditions, individuals are randomly assigned to treatments to ensure that the probability of receiving a treatment is not linked to the outcome of interest. Of course, many channels can affect a given outcome. Since the treatment likely impacts only one such channel, analysts must be careful that the treatment assignment is not confounded by any other channels. For example, an increase in a reenlistment bonus is one channel for increasing retention. If an imminent war affects both reenlistment rates and bonus levels, the causal link between the bonus and retention is confounded. Confounding variation must be accounted for in the analysis.

Outside of experimental conditions, researchers need to account for the possibility that treatment status is not random. For example, treatment status may be correlated with particular individual characteristics. If this selection is based on features that are recorded in the data, then researchers can control for those features.[27] Causal claims are possible if, after controlling for those features, selection into treatments is random.[28] Knowing what features to control for and knowing that no other features exist that need to be controlled for requires an underlying theoretical model. If the analyst effectively controls for all the features that affect the probability of being treated (in accordance with an appropriate model), then the estimation of the treatment's effect is *unconfounded*. That is, the effect of the treatment is directly measurable after controlling for confounding factors.

Another complication is that the effect of treatment cannot be observed for everyone in the data. Ideally, each individual would be observed in the data with and without the treatment. However, observational data contain some individuals who are treated and others who are not since one person cannot be treated and untreated at the same time. In *panel data* settings, where people or objects are repeatedly observed over a period of time, individuals are sometimes observed in treated and untreated states in different periods. Moreover, the distribution of individuals across the treated and untreated categories is often associated with the observable characteristics of the individuals. In this case, estimating the treatment effect requires grouping similar individuals and weighting these individuals with associated probabilities, typically through the use of propensity scores, so that the resulting effect estimates reflect a broader population.

A *propensity score* is the probability that an individual receives a treatment based on the characteristics of the individual that can be observed in the data. If an individual shares one or more key attributes with people who commonly receive a given treatment, then the individual will have a higher propensity score (i.e., a higher propensity to receive the treatment). Conversely, if an individual has attributes that differ from those who are commonly treated, then the individual will have a lower propensity score (i.e., a lower propensity to

[27] The selection based on features that are recorded in the data is referred to as "selection on observables."

[28] More formally, the selection into treatments must be independent of unobservable characteristics.

receive the treatment). The use of propensity scores for causal studies relies both on the unconfoundedness assumption and on an assumption that every individual has at least some positive probability of being assigned to each treatment. Rosenbaum and Rubin,[29] in their seminal work, found that under these assumptions, conditioning on propensity scores removed the bias from the observed variables in calculating average treatment effects. Moreover, by matching pairs of individuals with similar propensity scores (where one individual in the pair received the treatment and the other did not), it is possible to examine causal effects.[30]

While propensity scores often look at individuals observed at a single point in time, *difference-in-difference*s is a useful tool when individuals in a control and treatment group are observed repeatedly over time—both before and after the administration of a treatment. It critically assumes that the time trend of the outcome for individuals in the treated and untreated groups would have been the same if no treatment had been given. Thus, if the assumptions hold, changes to the time trend can be causally attributed to the policy. The method's name comes from the two sets of differences that are needed to highlight the change in the time trend. First, we take the before and after time difference within each group, which results in a time difference for the treated group and a time difference for the control group. Second, we take the difference between these two time differences to obtain the estimated treatment effect. The magnitude (and possibly the direction) of this estimated effect may be inaccurate if individuals in either the treatment or control group modified their behavior before the treatment in anticipation of the treatment since that would violate the critical assumption about the time trends.

Another key assumption of difference-in-differences (and other treatment effect estimation methods) is that the underlying characteristics of the treatment and control group are similar enough for a comparison to be relevant. Different time trends could be due to a number of factors. The impact of the treatment is clearest when the treatment and control are otherwise similar. In practice, it can be a challenge to find adequately similar controls. If a treatment occurs at the level of a city, what characteristics of that city need to be

[29] Paul R. Rosenbaum and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70, no. 1 (01 April 1983): 41–55, https://doi.org/10.1093/biomet/70.1.41.

[30] Austin provides an overview of the literature on propensity scores. Although adjusting for propensity scores removes bias in calculating the average treatment effect, the method often suffers from a loss of statistical efficiency—meaning that more observations are needed to achieve a given level of statistical performance (see Peter C. Austin, "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," *Multivariate Behavioral Research* 46, no. 3 (2011): 399–424, doi:10.1080/00273171.2011.568786). Hirano, Imbens, and Ridder address this problem by devising a non-parametric estimate for propensity scores that can be used to obtain unbiased and statistically efficient estimates of the treatment effect (see Keisuke Hirano, Guido W. Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71, no. 4 (July 2003): 1161–1189, https://doi.org/10.1111/1468-0262.00442.

matched to provide an adequate comparison (e.g., its population, crime rate, income distribution, demographics, climate, economy, or other dimensions)?[31] Analogously, if a policy is implemented on a military base or on a naval ship, identifying a comparison group that can serve as an appropriate baseline may not be entirely obvious. For instance, is Kadena Air Base in Japan adequately similar to Ramstein Air Base in Germany since both are major overseas bases, or is the much smaller Spangdahlem Air Base a better comparison to Ramstein Air Base since both are in Germany?

The *synthetic control* method adds a degree of flexibility for identifying a baseline in a case such as this. Instead of directly pairing a given treatment with a given control one-to-one, the treatment is matched to a weighted combination of potential controls. These weights are calibrated to highlight the dimensions along which the treatment group and each of the potential controls are most similar. For example, one base may be a better match in terms of region, but another base may be a better match in terms of its demographic composition. The resultant baseline is thus a hybrid, synthetically created from several potential controls. A benefit of creating such a synthetic baseline is that it clarifies the similarities and differences between the treated and control groups.[32]

Another tool for identifying causal effects focuses on boundary cases where there is a distinct change in policy that results from being on one side of a cutoff or threshold than the other. Cutoffs are common in many eligibility requirements, such as when public schools restrict enrollment to those living within rigid geographic boundaries, when participation in a government benefit has strict age or income requirements, or when individuals along state boundaries face different tax laws. If individuals have similar features on either side of the boundary and it is difficult to switch from one side to the other, then *regression discontinuity* can be used to estimate whether the boundary coincides with a jump in the outcome of interest. The magnitude of the jump represents the effect of moving from one side of the boundary to the other.[33] In the military, for instance, analysts could estimate the effect of the Blended Retirement System (BRS) on retention by looking at the

[31] The example of trying to find an adequate match for a city is illustrated in Card's analysis of a population shock to the Miami labor market. Lacking a clear comparison city, Card uses difference-in-differences to compare Miami with four potential comparison cities: Atlanta, Houston, Los Angeles, and Tampa/St. Petersburg. See David Card, "The Impact of the Mariel Boatlift on the Miami Labor Market," *Industrial and Labor Relations Review* 43, no. 2 (1990): 245–257, https://doi.org/10.1177/001979399004300205.

[32] See Alberto Abadie, Alexis Diamond, and Jens Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association* 105, no. 490 (2010): 494, https://doi.org/10.1198/jasa.2009.ap08746.

[33] Athey and Imbens provide an overview of regression discontinuity and other causal estimation techniques. See Susan Athey and Guido W. Imbens, "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives* 31, no. 2 (Spring 2017): 3–32, https://www.aeaweb.org/articles?id=10.1257/jep.31.2.3.

cohorts of individuals who were narrowly eligible to opt into the new BRS and those who were narrowly ineligible and were required to continue under the old retirement plan.

Service members with more than twelve years of service as of 31 December 2017 were ineligible to opt into the BRS. All other service members who entered military service before 2018 also had the option of opting into the BRS. A regression discontinuity design would compare individuals barely at twelve years of service to those barely below twelve years of service at the end of 2017. This effect is notably localized. It does not capture the overall change in the retention profile; rather, it captures only the change for those who were close to the cutoff. Still, this situation provides a valuable natural experiment for gaining a clear picture of the magnitude of the effect of the change in the retirement system for a cohort of otherwise similar service members.[34]

One of the oldest econometric techniques for examining causality departs from the statistical tradition of using randomized or quasi-randomized experiments. The premise of randomization is to ensure that the probability of being treated only affects the outcome through observable variables. The departure from the statistical tradition is to look at situations where all the factors that affect the probability of being treated cannot be observably accounted for—particularly when those factors also affect the outcome of interest. Analysis in such situations is *confounded* since a direct estimation of the treatment effect is not possible. In this setting, selection into a treatment may be driven by individual choice or by other factors within the individuals' environment. Contrasted with chance treatment assignments in experiments, Imbens notes that "[e]xposure to treatment is rarely solely a matter of chance or solely a matter of choice. Both aspects are important and help to understand when causal inferences are credible and when they are not."[35]

That said, identifying a treatment effect still requires unconfounded variation. *Instrumental variables* provide a method for identifying this unconfounded variation. An instrument is a variable that does not directly affect the outcome of interest but does affect the probability that an individual receives a treatment. When the instrument changes values,

[34] The retention impacts of the BRS are still too new to examine using this methodology. Sufficient data will be available in five to ten years. Another methodological factor to consider is that individuals near the cutoff only differed in terms of their ability to opt into the new retirement system. The ability to opt in is not as sharp a delineation as a situation in which service members with fewer than twelve years of service are all automatically placed under the new retirement system. Opting in allows for a positive probability of switching retirement systems, but the change is not certain. A variant of regression discontinuity can account to some degree for situations such as this one, where the delineation is "fuzzy." However, an analysis comparing those who opt in vs. those who are eligible to opt in but choose not to would likely require the use of instrumental variables.

[35] Guido W. Imbens, "Instrumental Variables: An Econometrician's Perspective," *Statistical Science* 29, no. 3 (August 2014): 324, https://www.jstor.org/stable/43288511.

the probability of receiving the treatment changes. Any change in the outcome can then be attributed to the change in the instrument or to the probability of treatment.[36]

In a classic example, Angrist[37] attempts to quantify the effect of serving in the military during the Vietnam era on individuals' long-term earnings. However, this effect cannot be directly calculated as the simple difference between the earnings of veterans and civilians. Some individuals, for instance, may enlist in the military because their earnings prospects in the civilian labor market are limited. If such individuals have lower earnings later as veterans, determining how much of the gap is directly attributable to military service and how much is attributable to individuals' underlying characteristics can be difficult.

Angrist disentangles the effect of military service by using random variation in the lotteries that determined the individuals' military draft priority. Those individuals with lower lottery numbers were more likely to enter military service (either by being drafted or by preempting the draft and voluntarily joining the military). Since the variation in the lottery numbers affected the probability of entering the military and since the lottery numbers were unrelated to long-term labor market earnings (except through military service), these numbers can be used as an instrument to identify the effect of military service in the Vietnam era on individuals' long-term earnings. Angrist finds that as of the early 1980s, military service for white veterans resulted in earnings that were about 15 percent less than non-veterans. This difference was equivalent to about two fewer years of experience in the civilian labor market, suggesting that time spent in military service was not a direct substitute to time spent in the civilian labor market (the median military service length in the Vietnam era was thirty-seven months). Military service did not result in a statistically significant earnings difference between nonwhite veterans and nonwhite civilians.

B. Machine Learning Contributions to Model Selection

Achieving valid estimates requires controlling for the appropriate set of confounding variables, but identifying those controls is a challenge. There may be many candidate controls, and many more candidate combinations. The full set of possible choices often becomes unwieldy. Critically, naïve model selection can invalidate estimates of the size and direction of treatment effects. A nascent literature examines how machine learning techniques can identify the appropriate set of controls while still enabling valid statistical

[36] The instrumental variables method is implemented in two stages. The first stage estimates the distribution of the treatment conditional on the instrument and any observable factors (beside the treatment) that may impact the outcome of interest. The second stage takes the estimated values for the treatment from the first stage and combines those values with observable factors to estimate the impact of the treatment on the outcome.

[37] Joshua D. Angrist, "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *The American Economic Review* 80, no. 3 (June 1990): 313–336, https://www.jstor.org/stable/2006669.

inference within the model. Machine learning can be used to determine which variables should enter the model, in which combinations, and with what weights.

The *partial linear model* provides a flexible regression setting for harnessing machine learning to identify key variable interactions. In this model, the outcome is linear in one or more key policy or treatment variables (just as in a *standard regression model*). However, the remaining variables, including any interactions or non-linear transformations of them, are modeled as an arbitrary function. Approximating this function is a prerequisite for estimating the treatment effect. This approximation includes dimension reduction, where the importance placed on variables and interactions that have minimal impact on the outcome is reduced or removed entirely. Dimension reduction does not come for free, however. As variables are removed, statistical bias can enter the model, requiring additional corrections to the estimates or structure to preserve key statistical properties.

Belloni, Chernozhukov, and Hansen[38] approach this setting with the assumption that the arbitrary function can be modeled by a linear combination of a relatively small number of control variables that are unknown to the researcher. They provide a simple procedure for selecting variables that maintains valid statistical inference. Building on an instrumental variables strategy, their method identifies the variables that directly predict the outcome variable and those that predict the outcome variable through the treatment variable. This approach captures variation that is useful in prediction but would be missed by focusing only on the variables that directly predict the outcome. Their method is as follows:

1. Identify the set of control variables that are predictive of the treatment variable.
2. Identify the set of control variables that are predictive of the outcome variable.
3. Estimate the effect of the treatment on the outcome by regressing the outcome on the treatment and the union of the variables identified in steps one and two.

The first two steps employ the machine learning technique of *regularization*, which is a process that penalizes model complexity. Belloni, Chernozhukov, and Hansen specifically focus on the Lasso method of regularization. Lasso, short for Least Absolute Shrinkage and Selection Operator, weighs the explanatory power of each control variable and only keeps variables that contribute substantially.[39] This procedure allows the researcher

[38] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen, "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies* 81, no. 2 (April 2014): 608–650, https://www.jstor.org/stable/43551575.

[39] The lasso method uses a penalty that forces to zero the coefficients on control variables that contribute minimally to the model—effectively removing them from the model. Ridge regression, on the other hand, uses a different penalty that dampens but does not eliminate the contribution of such control variables. Another regularization technique, elastic net, employs lasso and ridge penalties. Under certain conditions, there is an equivalence between elastic net, lasso, and support vector machines (a machine learning technique prominent in the early 2000s). See Quan Zhou et al., "A Reduction of the Elastic Net to Support Vector Machines with an Application to GPU Computing," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Palo Alto, CA: AAAI Press, 2015), 3210–3216,

to make statistically robust statements about the effect of the treatment on the outcome of interest.[40]

To achieve valid statistical inference in more general settings, Chernozhukov et al.[41] return to the partial linear model but with fewer underlying assumptions than those in Belloni, Chernozhukov, and Hansen[42]. Most notably, they do not require the arbitrary function in the partial linear model to be a linear combination of a relatively small number of control variables. Identifying a solution in this more general setting requires overcoming a fundamental problem: when machine learning methods are used to estimate the arbitrary function, it biases the estimate of the treatment parameter. The bias is a product of overfitting and of the bias that comes through the regularization process.[43] The solution to remove the bias incorporates the logic of instrumental variables. The outcome of interest is a function of a treatment variable and the numerous control variables, but the treatment variable itself may be a function of the control variables. The approach removes the effect of the control variables on the treatment variable, enabling valid inference on the objects of interest. Their *double machine learning* procedure is as follows:

1. Estimate the function relating the control variables to the outcome variable with machine learning.
2. Estimate the function relating the control variables to the treatment variable with machine learning.
3. Using the estimates from steps one and two, the authors construct an unbiased estimate for the treatment parameter.[44]

http://www.cs.cornell.edu/~kilian/papers/aaai15_sven.pdf; Martin Jaggi, "An Equivalence between the Lasso and Support Vector Machines," in *Regularization, Optimization, Kernels, and Support Vector Machines*, ed. Johan A. K. Suykens, Marco Signoretto, and Andreas Argyriou (Boca Raton, FL: CRC Press, 2015), 1–26.

[40] Formally, Lasso regularization produces a statistically consistent estimator and valid confidence intervals. Consistency is a statistical property implying that as the number of observations grow, the estimate converges to the underlying true parameter governing the data generating process.

[41] Victor Chernozhukov et al., "Double/Debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal* 21, no. 1 (February 2018): C1–C68, https://doi.org/10.1111/ectj.12097.

[42] Belloni, Chernozhukov, and Hansen, "Inference on Treatment Effects."

[43] Since regularization penalizes model complexity, it removes or dampens the impact of variables that add to model complexity, even if they may be relevant to the data-generating process in describing the relationship between the treatment variable and the outcome. Removing relevant variables from regression models, even if the impact of those variables is relatively small, biases the estimates of the parameters remaining in the model.

[44] To obtain not only an unbiased estimator, but also an efficient one, the authors divide the sample in two. They perform the two machine learning estimates with one half of the sample and then use those estimates to construct the treatment effect from the other half of the sample. They then switch the role of the two samples and average the two estimates of the treatment effect. This cross-fitting produces an efficient estimator.

Chernozhukov et al.'s[45] method produces unbiased, normally distributed point estimates with valid confidence intervals. It also has the flexibility to accommodate a broad set of machine learning methods in the estimation stages, including Lasso, random forests, gradient boosted forests, and neural networks. Double machine learning is not limited to model selection and can be used when there is an endogenous treatment variable that would typically invite an instrumental variable approach.

C. Experiments and Causal Forests

Experiments are crucial for capturing causal effects that may be difficult or impossible to identify from naturally occurring data. Consequently, experiments are standard practice in evaluating medical techniques and pharmaceutical products. They are a driving force in online marketing, where major corporations randomly show users slightly different versions of their website to see which version results in greater sales (a practice known as A/B testing). Experimentation is increasingly used to identify the efficacy of assistance programs in helping individuals rise above extreme poverty.[46]

In the defense context, experiments were actively used in DOD in the years immediately following the creation of the All-Volunteer Force. The sustainability of a military system without conscription was under high levels of scrutiny from Congress and Pentagon leadership, which resulted in deliberate experiments to evaluate policy efficacy. For instance, the military's ROI for bonuses and for offering contracts with varying commitment lengths was not well understood. During this early era of personnel policy experiments, the Army conducted a national experiment from 1982 to 1984 during which recruits in 70 percent of the nation were offered a $5,000 bonus for a four-year enlistment, recruits in 15 percent of the nation were offered $8,000 for a four-year enlistment, and recruits in the remaining 15 percent of the nation were offered the choice of $8,000 for a four-year enlistment or $4,000 for a three-year enlistment. The higher bonuses were restricted to recruits with higher test scores. The experiment examined the effect of the bonus on the overall number of recruits, the quality of the recruits, and the duration of enlistments for recruits. Compared to the $5,000 baseline, the two alternate treatments increased the overall number of recruits by 4 to 5 percent, increased the number of high-quality recruits in skill sets eligible for the bonus by 32 to 42 percent, and increased the overall number of person-years obligated to the Army by 6 to 8 percent.[47]

[45] Chernozhukov et al., "Double/Debiased Machine Learning."

[46] The 2019 Nobel Prize in economics was awarded to Abhijit Banerjee, Esther Duflo, and Michael Kremer for their work in establishing experimentation as a way to quantify the impact of assistance programs for those in extreme poverty. By quantifying the impact of these programs, it is then possible to focus assistance on programs that are cost effective and yield high returns.

[47] See J. Michael Polich, James N. Dertouzos, and S. James Press, *The Enlistment Bonus Experiment*. R-3353-FMP (Santa Monica, CA: RAND Corporation, 1986), https://www.rand.org/pubs/reports/R3353.html.

Replicability is a foundational scientific principle for experiments. To fully separate the research stages of hypothesis generation and hypothesis testing, many experiments follow the practice of formally registering the details of the experiment before it begins. Registration involves specifying how data will be collected, modeled, and tested.[48] This added rigor minimizes the occurrence of false positive results. A drawback is that it can limit the exploration and identification of unanticipated results, such as identifying subgroups for whom the treatment results in genuinely different effects. If the subgroups were not identified in the registered hypotheses, the results could be dismissed as spurious.

Wager and Athey[49] address the need to maintain statistical credibility in results that are published and propagated while also allowing for a data-driven solution to uncover treatment effects that differ across subgroups in unanticipated ways. Their application of random forests can identify heterogeneous treatment effects within subgroups while maintaining the statistical properties needed for valid inference. Specifically, their *causal forests* method produces unbiased, consistent estimates with valid confidence intervals. It relies on the unconfoundedness assumption mentioned earlier (which often holds more readily with data from experiments than with observational data). The following is a sketch of the causal forests method:

1. Take a random sample of observations of the training set and divide the sample in half. Dividing the sample prevents double-dipping from the same information when building a tree and later when making inference about that tree.
2. Use one half of the random sample to grow a tree. At this stage, all the data from one half of the sample can be used to determine the splits in the tree.
3. Use the second half of the training data—that not used in step 2—to estimate the treatment effect for each leaf of the tree. Each leaf must have at least a minimum number of observations from the treatment and control groups.
4. The preceding three steps create a single *causal tree*. Repeating these steps a sufficient number of times creates a *causal forest*. The treatment effect for the forest comes from averaging the treatment effects from each tree.

D. Matrix Completion

Some prediction exercises can be represented as a matrix of rows and columns where some entries are present, others are missing, and the goal is to deduce meaningful values for the missing entries. For example, if each row represents a service member, then each

[48] For instance, the Center for Open Science (https://cos.io/prereg/) allows analysts to register their research plans.

[49] Stefan Wager and Susan Athey, “Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests,” *Journal of the American Statistical Association* 113, no. 523 (2018): 1228–1242, https://doi.org/10.1080/01621459.2017.1319839.

column can represent a point in time and the values can be a measure of each service members' deployability at each point in time under a new training regime. Another matrix similarly tracks each service member's deployability at each point in time under an old training regime. The ideal scenario would be to compare each service member's readiness at each point in time under the new and old training regimes. The problem, however, is that the service member only experiences one training regime or the other at any given point in time. Consequently, each matrix will have many missing values. The goal then is to determine those missing values.

Recent algorithmic and computational advances have made matrix completion for pure prediction feasible.[50] Athey et al.[51] adapted these breakthroughs to conduct causal analysis on problems that involve panel data[52] by observing that the matrix completion problem can encompass several common causal analysis settings. For instance, in the new training regime example, the new training could be implemented in a number of different ways. If the new training regime only occurs during a single time period and if it is only given to some service members, then the problem resembles a cross-sectional analysis where there is a large treatment group, large control group, and no time dimension to the study—the typical setting for policy analysis under the unconfoundedness assumption. If, instead, a small number of service members experience the new training regime for several time periods while the remaining service members continue to use the old training regime, then time becomes a meaningful dimension. The question is not just concerned with the static difference of changing training regimes, but also with the trajectory from such a change over time. This setting is analogous to the synthetic controls literature. If the new training regime is rolled out gradually so that some service members begin using it in earlier periods and others begin using it in later periods, then the problem corresponds to the literature on fixed effects.[53]

[50] For instance, the 2006 Netflix competition for improving its algorithm for recommending films to customers is a matrix completion problem. The rows represent customers, the columns represent films, and the values represent customer ratings of films they have seen. The problem is to predict customer ratings for films they have not seen. Mazumder et al. present a computationally efficient and scalable method for predicting missing values in large matrices (such as the 2006 Netflix Prize data with its half a million customers and 18,000 films) (see Rahul Mazumder, Trevor Hastie, and Robert Tibshirani, "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *Journal of Machine Learning Research* 11 (2010): 2287–2322, https://dl.acm.org/doi/10.5555/1756006.1859931). Their SOFT-IMPUTE algorithm is the workhorse in the Athey et al. paper (in 2018, see following footnote) for applying matrix completion to problems of causal analysis.

[51] Susan Athey et al., *Matrix Completion Methods for Causal Panel Data Models*, NBER Working Paper No. 25132 (Cambridge, MA: National Bureau of Economic Research, October 2018), https://www.nber.org/papers/w25132.pdf.

[52] In panel data, individuals are observed across many time periods.

[53] Fixed effects can statistically account for unobserved individual characteristics that remain constant over time and for unobserved time-specific effects that are constant across individuals. In traditional models,

Matrix completion also provides a means for estimating alternative, unobserved outcomes. Athey et al. specifically focus on estimating the missing values in the matrix of untreated observations so that they can compare the untreated state to the treated state. Moreover, compared to related tools for addressing similar causal problems (e.g., synthetic controls and difference-in-differences), their prediction errors are low.[54] However, the matrix completion method does not currently permit statistical inference. It imputes the missing values for the untreated state but does not allow formal statistical comparisons between the treated and untreated states.

E. Deep Instrumental Variables

The use of instrumental variables borders between the *how* (or *why*) questions and the *what if* questions. Estimating treatment effects of retrospective policies, such as the effect of Vietnam era military service on long-term earnings, falls squarely on the side of *how* (or *why*). In other cases, instrumental variables can be used to correctly model the underlying structural relationship between variables. A classic example is modeling the interplay between supply and demand. Economic data are often limited to information about price and quantity. Changes in either could signify a shift in the supply curve, in the demand curve, or in both. However, understanding the shape (or *elasticity*) of each curve requires moving one while holding the other fixed, and that cannot be done with price and quantity information alone. Price and quantity pinpoint the intersection of these two curves but generally do not trace the curves out. Plotting each distinct curve requires an instrumental variable.[55] Understanding the shape of the supply and demand curves enables the analysis of some *what if* questions, such as what would happen if a tax was implemented or if a tariff was removed.

Supply and demand curves exist in the military context. For example, the price may be a recruiting bonus, and the quantity may be the number of recruits accessed. Instrumental variables are required to understand the supply of potential recruits, which enables calculation of how the number of recruits would change with different recruiting bonuses.

the individual effects and the time effects are treated as additive. More recent models, known as *interactive fixed effects,* permit a multiplicative relationship between the individual and time effects.

[54] Although this is not a formal mathematical result, Athey et al. provide empirical examples comparing prediction errors across methods. See Athey et al., *Matrix Completion Methods,* 25–28.

[55] Instrumental variables were introduced by Philip Wright in 1928 specifically as a way to isolate information about supply and demand curves. “In the absence of intimate knowledge of demand and supply conditions, statistical methods for imputing fixity to one of the curves while the other changes its position must be based on the introduction of additional factors. Such additional factors may be factors which (A) affect demand conditions without affecting cost conditions ….” See Philip G. Wright, *The Tariff on Animal and Vegetable Oils* (New York, NY: Macmillan Company, 1928), 311–312.

Hartford et al.[56] have adapted instrumental variables for use within deep learning neural network frameworks. They specifically use neural networks to develop a more flexible way for modeling the underlying relationship between the variables in the model. The goal is to correctly capture these relationships and thus enable predictions for what the outcome would be under alternative treatments. However, their method only generates predicted outcomes under alternative treatments. It does not permit a statistical comparison of outcomes under different treatments (similar to the matrix completion method in Athey et al.[57]).

The analysis in Hartford et al.[58] goes to the underlying mathematical formulation for instrumental variables. It is common to assume that the relationships being modeled in each stage of the instrumental variable estimation procedure are linear; however, they also allow for complex, interconnected relationships. They use neural networks in each stage of the instrumental variables estimation procedure. The first stage uses standard off-the-shelf software for the neural network computation. The second stage, however, uses a neural network with a novel loss function.[59] Researchers should expect advances in the coming years to build on this impressive start.

[56] Jason Hartford et al., "Deep IV: A Flexible Approach for Counterfactual Prediction," in *Proceedings of the 34th International Conference on Machine Learning* – Volume 70, ed. Doina Precup and Yee Whye Teh (JMLR.org, 2017), 1414–1423, http://proceedings.mlr.press/v70/hartford17a.html.

[57] Athey et al., *Matrix Completion Methods*.

[58] Hartford et al., "Deep IV: A Flexible Approach."

[59] Directly calculating the second stage estimates is a substantial computational burden. However, the authors find that estimating an upper bound of the loss function (as defined by Jensen's inequality) instead of the loss function itself produces results of a similar quality and is less computationally demanding.

4. “What if?” Questions: Prospective Policy Analysis

If we are not to get lost in the overwhelming bewildering mass of statistical data that are now becoming available, we need the guidance and help of a powerful theoretical framework.[60]

– Ragnar Frisch, 1933

When assessing the potential impacts of a policy that has not yet been implemented, a theoretical framework is required. Without the benefit of hindsight to reveal actual effects of the policy (even among a small or loosely related group), we must turn to first principles, which requires identifying a core set of rules or guiding assumptions that adequately represent the human decision-making process. These rules allow us to hypothesize how people (or systems) would behave if offered different conditions or choices. Since human decision making is often irrational and inconsistent, describing behavior using rigid mathematical structures introduces the possibility of mistaken conclusions. Yet, by adopting such rules, we can begin to estimate the impact of a new policy. These rules can also provide “intuitive insights into the decision making process …”[61] and enable the model to “simulate behavior under new circumstances ….”[62]

Major career decisions, such as whether to renew a service contract, often have immediate and long-term implications. Choices made today impact the range of choices available tomorrow. The dynamic interplay between an individual’s current and future valuation of pursuing one career trajectory vs. another has often been modeled with the mathematical apparatus of *dynamic programming*.[63] At its core, dynamic programming is a

[60] At nearly a century old, this quote is a reminder that the challenge of grappling with “big data” is not a new phenomenon. Ragnar Frisch was an early pioneer in the field of econometrics and served as the first editor of the now reputable journal *Econometrica*. When the Nobel Prize in Economics was introduced in 1969, he and Jan Tinbergen were recognized as its first recipients. This quote is from the editor’s note that Frisch penned for the inaugural issue of *Econometrica*. See Ragnar Frisch, “Editorial,” *Econometrica* 1, no. 1 (January 1933): 2, https://www.jstor.org/stable/1912224.

[61] Stefano Ermon et al., “Learning Large-Scale Dynamic Discrete Choice Models of Spatio-Temporal Preferences with Application to Migratory Pastoralism in East Africa,” in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (Palo Alto, CA: AAAI Press, 2015), 645, https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9725/9308.

[62] Ibid., 648.

[63] Richard Bellman coined the phrase “dynamic programming” in the 1950s. In an environment that he felt was hostile to mathematical research, he chose the name to hide “the fact that [he] was really doing

mathematical optimization exercise in which an individual makes choices in an attempt to maximize her well-being now and in the future. Assuming certain technical conditions hold, this multi-period optimization problem can be rewritten as a recursive two-period problem. In other words, the individual balances the tension between fulfilling the demands and desires of today vs. fulfilling the demands and desires of tomorrow. Tomorrow, she faces the same tension between that day and the next. Thus, each period can be viewed as a two-period problem that has a subsequent two-period problem buried inside of it. If each two-period problem is identical, then the optimal choice-balancing in one two-period problem is the optimal choice for all two-period problems.

If there is a clear terminal period that differs from the other periods (which can occur in finite cases), then after solving the final two-period problem, the penultimate problem can be solved and so on all the way back to the first two-period problem. However, the complicating factor is that this decision structure grows like a tree, with the first two-period problem as the trunk and subsequent two-period problems as branches that fork depending on the decisions made in the previous periods.[64] Thus, there is not just one final two-period problem. Depending on the number of periods and the number of choices at each period, there may be countless end states. Every possible path need not be calculated. Some paths will be inferior. By starting at the end and working backward, the *backward recursion* algorithm prunes away the inferior paths and focuses on the promising ones.

As early as 1979, Glenn Gotz and John McCall of RAND began applying dynamic programming to the problem of examining "the incentives to retire under alternative retirement systems."[65] Building on the work of Heckman and Willis,[66] Gotz and McCall subsequently created the Dynamic Econometric Retention Model (DERM) in 1980. By 1984, they had renamed it the Dynamic Retention Model (DRM). This model has been used

mathematics inside the RAND Corporation." It was a name "that not even a Congressman could object to." See Richard Bellman, *Eye of the Hurricane: An Autobiography* (Hackensack, NJ: World Scientific Publishing Co., Inc., 1984), 159.

[64] In finite horizon dynamic programming problems, branches may also come back together. Another common analogy is to consider the problem of finding the shortest possible path from one point in a city to another. Going three blocks north and two blocks west arrives at the same point as going one block west, two blocks north, one block west, and one block north. However, depending on traffic patterns, one route may be faster than the other.

[65] Glenn A. Gotz and John J. McCall, *A Sequential Analysis of the Air Force Officer's Retirement Decision*, N-1013-1-AF (Santa Monica, CA: RAND Corporation, 1979), v, https://www.rand.org/pubs/notes/N1013-1.html.

[66] James J. Heckman and Robert J. Willis, "Estimation of a Stochastic Model of Reproduction: An Econometric Approach," in *Household Production and Consumption*, ed. Nestor E. Terleckyj (Cambridge, MA: National Bureau of Economic Research (NBER), 1976), 99–146, https://www.nber.org/books/terl76-1.

widely since for estimating “voluntary retention rates under a broad range of [proposed] compensation, retirement, and personnel policies.”[67]

Gotz and McCall’s early work was pioneering and was cited as one of “the first examples of estimable econometric models which are derived from discrete stochastic control problems which do not have closed-form solutions.”[68] Yet, despite their breakthrough, their modeling framework was computationally limited due to the burdensome process of solving numeric dynamic programming problems. This limitation was not unique to their work but permeated related models of dynamic discrete choice environments. The prevalent technique entailed computing “the valuation function using backwards recursion, not just once, but every time the parameters [were] evaluated in the estimation routine.”[69] This computational complexity curtailed the extent to which nuances of an environment could be modeled, a consequence known among analysts as the “curse of dimensionality.” For military retention, the model was practically reduced to estimating the human decision-making process based on monetary compensation alone, and, today, that is still largely the case. Excursions to extend the DRM to take into account other potential factors have made only incremental progress, leaving many multi-faceted aspects of the decision making process unmodeled.[70]

Academic advances have attempted to devise modeling strategies and algorithms that confront this computational challenge. Hotz and Miller[71] proposed a method to sidestep the problem of explicitly solving the valuation function at every step of the backward recursion routine. They note that many dynamic choice problems have a terminal choice that prevents subsequent choices. In the military retention context, exiting the military is a terminal

[67] Glenn A. Gotz and John J. McCall, *A Dynamic Retention Model for Air Force Officers: Theory and Estimates,* R-3028-AF (Santa Monica, CA: RAND Corporation, December 1984), v, https://www.rand.org/pubs/reports/R3028.html; Glenn A. Gotz and John J. McCall, *Estimating Military Personnel Retention Rates: Theory and Statistical Method*, R-2541-AF (Santa Monica, CA: RAND Corporation, June 1980), https://www.rand.org/pubs/reports/R2541.html.

[68] The quote is from John Rust’s influential 1987 paper that was recognized as the best publication in *Econometrica* within a five-year period by its receipt of the Frisch Medal. See John Rust, “Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher,” *Econometrica* 55, no. 5 (September 1987): 1014 (footnote 10), https://www.jstor.org/stable/1911259.

[69] V. Joseph Hotz and Robert A. Miller, “Conditional Choice Probabilities and the Estimation of Dynamic Models,” *The Review of Economic Studies* 60, no. 3 (1993): 498, https://www.jstor.org/stable/2298122.

[70] For example, Asch et al. extend the DRM to incorporate the cohort of which an individual is a part at the time a policy change occurs (cohort is defined in terms of years of service). Cohort labeling allows the DRM to not just model the final steady state following the policy change, but also the transition to the policy change. Like the DRM of the 1980s, the decision to stay or leave is still based on a coarse, two-dimensional approximation of the military experience: monetary compensation and a catch-all taste for military service. See Beth J. Asch, Michael G. Mattock, and James Hosek, *A New Tool for Assessing Workforce Management Policies over Time: Extending the Dynamic Retention Model*, RR-113-OSD (Santa Monica, CA: RAND Corporation, 2013), https://www.rand.org/pubs/research_reports/RR113.html.

[71] Hotz and Miller, “Conditional Choice Probabilities.”

choice since it precludes pursuing the various career trajectories that may be available from remaining within the military. This terminal choice can serve as a reference point from which the valuation of the other trajectories can be measured. Moreover, these valuations can be expressed—not directly, but in terms of the "probabilities that particular choices occur, given the observed state variables,"[72] the transition probabilities across states, and expected payoffs. The choice probabilities and the transition probabilities are relatively easy to obtain. Once in hand, they can be used to estimate expected payoffs. Payoffs are in terms of utility or individual preferences.[73]

Running in parallel to the development of structural models of dynamic discrete choice problems has been the development of *inverse reinforcement learning* in the sphere of machine learning. These two fields developed largely independently toward the same goal: to infer the underlying objective and reward system that motivates observed behavior. As Ng and Russell[74] said, "[i]n examining animal and human behavior we must consider the reward function as an unknown to be ascertained through empirical investigation. … Consider, for example, that the bee might weigh nectar ingestion against flight distance, time, and risk from wind and predators,"[75] but how can the researcher determine the relative weights in the bee's risk and reward function? Analogously, in modeling military service members, the researcher needs a methodology for determining the relative weights for training and skill development, travel experiences, combat experiences, compensation, assignment location, frequency of assignment changes, spousal and family commitments, and outside employment options for the service member and any family members. Stated differently, the researchers must specify a utility function for service members. Although the relative weights are not directly observable, the choices that individuals make over time may be. The sequence of circumstances (or states) under which choices are made may reflect in some way the consequences or rewards of previous choices—providing insight into the individual's underlying reward system.

[72] Peter Arcidiacono and Paul B. Ellickson, "Practical Methods for Estimation of Dynamic Discrete Choice Models," *Annual Review of Economics.* 3, no. 1 (2011): 365, https://www.annualreviews.org/doi/abs/10.1146/annurev-economics-111809-125038). See also Hotz and Miller, "Conditional Choice Probabilities," 501.

[73] Recent models of dynamic discrete choice for retention within the military have incorporated Hotz and Miller's approach of using conditional choice probabilities. See Colin M. Doyle, *The Military Career Analysis Model: Theory, Methodology, Estimation, and Application of a Unified Model of Military Personnel Accession, Retention, Promotion, and Life-Cycle Cost*, IDA Document P-5264 (Alexandria, VA: Institute for Defense Analyses, June 2015); Jared Huff et al., *Estimating the Retention Effects of Continuation Pay* (Arlington, VA: Center for Naval Analyses (CNA), April 2018), https://www.cna.org/CNA_files/PDF/DRM-2018-U-017177-Final.pdf.

[74] Andrew Y. Ng and Stuart J. Russell, "Algorithms for Inverse Reinforcement Learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ed. Pat Langley (San Francisco, CA: Morgan Kaufmann Publishers, Inc. 2000), 663–670, https://ai.stanford.edu/~ang/papers/icml00-irl.pdf.

[75] Ibid., 1.

Maximum entropy inverse reinforcement learning allows for the possibility that each individual's sequence of choices may not necessarily be optimal. Although individuals may attempt to optimize their welfare (or utility), there are mistakes, noise, and other variation. However, the method does assume that sequences of choices and states that result in higher welfare are more likely to be observed than sequences resulting in lower ones.[76] The mathematics underlying this assumption about the data-generating process has a logit structure, just like the logit structure that frequently appears in dynamic discrete choice modeling. The structure is similar enough that Ermon et al.[77] are able to assert an equivalence between the two methods when individuals do not discount the future.[78] Since dynamic discrete choice modeling permits discounting, it is more flexible. After making this connection, they refine the traditional algorithm for computing dynamic discrete choice models to make it more adaptable to large state spaces.

Unlike the Hotz and Miller[79] approach of sidestepping the dynamic programming problem, Ermon et al. return to the original algorithm to improve its speed and scalability. They incorporate a stochastic gradient descent into the algorithm to accelerate convergence. To identify the underlying reward system generating behavior over a fixed time horizon, the original algorithm takes the observed behavior in the final period and then determines what the reward system must look like at the second-to-last period to generate the last period results. It then proceeds to each previous period in turn, conducting calculations over the various possible actions and states for each individual in the data. Calculating all potential decision branches is a computationally difficult task that grows exponentially in the number of possible actions and states. The traditional algorithm evaluates the entire data set at each iteration. Rather than calculating the full dynamic programming matrix at each time period, only to discard the calculations and begin anew for the next time period, Ermon et al. update the main learning parameter after each period and use that information to fill the next column of the dynamic programming matrix. Updating in this

[76] In introducing maximum entropy inverse reinforcement learning, Ziebart et al. used an example of trying to uncover the utility function of taxi drivers. Their numerical estimation used extensive data on taxicabs in Pittsburgh, Pennsylvania, which was characterized into over 300,000 road segments (states) and 900,000 transitions at intersections (choices). Their work on maximum entropy inverse reinforcement has been cited in several robotics applications. See Brian D. Ziebart et al., "Maximum Entropy Inverse Reinforcement Learning," in *Proceedings of the 23rd National Conference on Artificial Intelligence* – Volume 3, ed. Anthony Cohn (Palo Alto, CA: AAAI Press, 2008), 1433–1438.

[77] Ermon et al., "Learning Large-Scale Dynamic Discrete Choice Models."

[78] Sharma Kitani, and Groeger make a similar statement. In a dynamic discrete choice model, "the softmax [or logit style] recursion is a consequence of Bellman's optimality principle," while under maximum entropy inverse reinforcement learning, the recursion arises "from an information-theoretic perspective." Sharma Kitani, and Groeger combine maximum entropy inverse reinforcement learning with the Hotz and Miller conditional choice probabilities. See Mohit Sharma, Kris M. Kitani, and Joachim Groeger, "Inverse Reinforcement Learning with Conditional Choice Probabilities," arXiv:1709.07597v1 (2017), 2.

[79] Hotz and Miller. "Conditional Choice Probabilities."

manner speeds computational time significantly and presents a scalable method for addressing complex state spaces.[80]

Norets[81] uses a similar logic to break down the traditional algorithm into pieces that can be approximated—rather than fully solved—at each iteration. He focuses on approximating the solution to the dynamic programming problem using feed forward neural networks. He mathematically demonstrates conditions under which the neural network approximation converge to actual solution.[82]

Identifying ways to cope with the curse of dimensionality is an ongoing line of research. A breakthrough by Scheidegger and Bilionis[83] provides a method for implementing dynamic programming on problems with high-dimensional state spaces. This method combines a supervised machine learning method known as *Gaussian process regression* with the recently developed *active subspace method*. The active subspace method identifies the portions of the state space where changes in the input values have the greatest impact on the outcome.[84] Gaussian process regression seeks to identify the mapping from a multivariate input space to a univariate output space.[85] It can handle dynamic programming problems with fewer than twenty (or so) dimensions—which is a significant

[80] The entire matrix is gradually updated over the course of several iterations. The gradient of the matrix is approximated in each iteration using stochastic gradient descent.

[81] Andriy Norets, "Estimation of Dynamic Discrete Choice Models Using Artificial Neural Network Approximations," *Econometric Reviews* 31, no. 1 (2012): 84–106, https://doi.org/10.1080/07474938.2011.607089.

[82] Norets is not the first to incorporate neural networks to some aspect of solving dynamic programming problems. In providing an overview of a handful of early attempts, Rust concluded that despite the benefits of neural nets, "the Achilles heel is the curse of dimensionality associated with solving the global minimization problem …." Neural nets "have many local minima" and finding one that "yields good approximations to the value or policy function can be extremely burdensome" (see John Rust, "Numerical Dynamic Programming in Economics," in *Handbook of Computational Economics*, Volume 1, ed. Hans M. Amman, David A. Kendrick, and John Rust (Amsterdam, The Netherlands: Elsevier Science B.V., 1996), 690–691, https://doi.org/10.1016/S1574-0021(96)01016-7). Norets cautions that his theoretical results "require finding the global minimum" but notes that in his experiments there were "no cases of getting stuck in a very bad local minimum" (see Norets, "Estimation of Dynamic Discrete Choice Models," 91).

[83] Simon Scheidegger and Ilias Bilionis, "Machine Learning for High-Dimensional Dynamic Stochastic Economies," *Journal of Computational Science* 33 (April 2019): 68–82, https://doi.org/10.1016/j.jocs.2019.03.004.

[84] The active subspace method is similar to principal component analysis (PCA) but with a different optimization objective. Each seeks to reduce the dimensionality of the input space. However, PCA identifies the linear projection of input space that best represents that *input* space, while the active subspace method identifies a linear projection of the input space that represents the *output* space.

[85] Following Bayesian processes, the method requires specifying priors for the mean and covariance for the function to be identified. It also requires explicit assumptions about the data-generating process for the training data (e.g., observations are independently drawn, with observed target values distributed around the true functional value according to some distribution). Based on these assumptions, the method can assess the probability of the values in the training data, and the priors can be updated from these probabilities.

achievement by itself. However, when combined with the active subspace method, it can effectively address dynamic programming problems with hundreds of dimensions.[86] A benefit of a model that can accommodate high-dimensional state spaces is that uncertainty within the model can be estimated by adding "the parameters that carry uncertainty as additional, continuous pseudo-states to the model."[87]

Collectively, these and other methodological advances offer significant promise for expanding the scope of *what if* questions that can be examined in defense personnel analysis. Dynamic programming has been at the center of these analyses for a generation. Machine learning techniques offer the potential to dramatically expand the state space used within dynamic programming problems to better capture the intricate and complex dimensions of military service. Still, machine learning advances in the sphere of *what if* questions remain less mature than for the *what* or *how* classes of questions. At present, there are no off-the-shelf algorithms for adaptation to specific *what if* question use cases, although progress to date illuminates how this might be achieved. Investment is needed to develop these tools for answering *what if* defense personnel policy questions. Just as the DOD invested in developing the basic science needed to apply dynamic programming to defense personnel policy questions through Gotz and McCall's *A Dynamic Retention Model for Air Force Officers* in the 1970s and 1980s, so the DOD can now lead a new generation of analytic progress by maturing the application of machine learning advances to structural policy analyses and modeling.

[86] Gaussian process regression, whether or not it is combined with the active subspace method, has an additional computational advantage of being able to approximate functions on state spaces with non-traditional geometries (one example being the simplex, which can be used to represent budgetary constraints and the tradeoffs between different combinations of goods and services).

[87] Scheidegger and Bilionis, "Machine Learning for High-Dimensional Dynamic," 77. The authors demonstrate their methodology in a macroeconomic model of an economy: a dynamic stochastic optimal growth model over infinite time. This framework is notably different from the finite-horizon, dynamic discrete choice setting that is common in defense personnel analysis.

This page is intentionally blank.

5. Concluding Recommendations

A. Enhancing Data Access, Quality, and Scope

Data are the lifeblood for advanced defense personnel research. Machine learning algorithms harness data by capturing complex patterns within these data. However, the degree to which the algorithms can capture informative patterns is based on the quality, coverage, and scope of the data. To better harness the information within its personnel data, DOD must make sustained investments to ensure that high-quality data are broadly accessible for research and development.

1. Data Access

Recommendation 1: *Reduce barriers to establishing data access and sharing across organizations within DOD and between DOD and other entities.*

Recommendation 2: *Harmonize data to enable meaningful linking of information across organizations.*

Recommendation 3: *Establish secure analytic computing environments where analysts in DOD and those who support DOD can access data, work, and collaborate effectively.*

As discussed in Chapter 2, DOD's personnel data assets are dispersed across numerous organizations within the DOD enterprise. Personnel data are frequently siloed within these separate organizations, and protocols for accessing and bridging data across organizations are often non-existent or difficult to navigate. Many of these individual organizations (e.g., DMDC or the personnel directorates of the military services) maintain data repositories that are extensive enough to successfully employ the machine learning tools that are discussed in this paper. However, the scope and quality of analysis will be limited unless bridges can be built to better enable data access for secure research purposes across organizational barriers. Such bridges take two major forms: policies for establishing data access across organizations and data harmonization to enable the data from the two organizations to be linked in a meaningful manner. The former is a strategic governance bridge for establishing permissions and access protocols, and the latter is a tactical bridge for identifying how the data relate to and complement each other.

Similarly, DOD should further the process of building data access agreements with other government organizations. For instance, coupling DOD's personnel records with records from the Department of the Treasury could help DOD to better understand the career paths of high-performing individuals after they exit the military. Such information could aid DOD in shaping retention and career management policies for its top performers.

Data access also requires a secure, flexible environment for hosting the data. The Office of the Under Secretary of Defense for Personnel and Readiness (OUSD(P&R)) has partnered with IDA in conceptualizing a data environment that can host a broad corpus of DOD personnel data and be a staging ground for linking personnel data across organizations.[88] Gleaning the breadth of insights from these data that can be used to enhance the management, lethality, and efficiency of the force requires many hands. DOD should make sustained investments in developing such a secure environment to support a scalable number of analysts throughout DOD, its Federally Funded Research and Development Centers, and other supporting research organizations.

2. Data Quality

Recommendation 4: *Create and implement a pipeline for transforming frequently used "raw data" items into "research ready" cleaned, standardized, documented data objects in a reproducible manner.*

Personnel data are frequently messy or incomplete, and require standardization, formatting, alignment, and definition before they can be used for analysis. Specific data cleaning tasks include standardizing data formats, normalizing relationships or units between data elements, ensuring consistent coding of data values, and distinguishing different forms of omitted data (e.g., data that were not collected because these data were not applicable vs. data that were applicable but were missing). These tasks can be time intensive. Maintaining a pipeline for transforming "raw data" into "research ready data" has been a significant and costly hurdle that has limited the accessibility and comparability of defense personnel data. The process of cleaning, standardizing, and harmonizing data

[88] See Julie Pechacek et al., *Considerations for Implementing a Defense Personnel Research Environment,* IDA Document P-9254 (Alexandria, VA: Institute for Defense Analyses, September 2018); Julie Pechacek et al., *User Requirements for the Enterprise Data to Decisions Information Environment,* IDA Document NS D-9139 (Alexandria, VA: Institute for Defense Analyses, August 2018).

should be fully documented in a reproducible (and, ideally, automated) manner. The researcher must also be able to understand what the data represent. Data dictionaries, metadata, and other documentation that captures the context, coverage, and properties of the data must be maintained. DOD, in building and maturing its personnel data as a long-term information asset, should ensure that data curation efforts are undertaken and sustained.

3. Data Scope

Recommendation 5: *Ensure that data are preserved over long periods of time to allow for studies of the entire relevant process.*

Recommendation 6: *Collect and preserve information on all choices and benefits offered to individuals—not just information on what option(s) an individual ultimately chooses.*

The scope of personnel data can be measured in terms of many dimensions, including longitudinal time coverage, coverage across major military populations (e.g., active duty, reserve, National Guard, retiree, dependent, civilian), the fraction of individuals in a given population who are represented, the number of characteristics recorded about each individual, and coverage across data maintained by separate DOD organizations. Each dimension is valuable for different forms of analysis. For instance, studies on life-cycle military career decisions are enhanced by longitudinal data because careers can be decades in the making and because longitudinal data captures different economic and wartime stresses on the DOD's ability to maintain its personnel. As a benchmark, it is helpful to have a record of personnel information during times of relative peace, high wartime operating tempos, economic prosperity, and economic downturns. Efforts to curate personnel data should seek to expand (rather than curtail) longitudinal time coverage and the scope of the other dimensions of personnel data.

Examining the effectiveness of various personnel policies requires information on people who received a given benefit and people who were eligible for a benefit but elected not to receive it. For instance, if a reenlistment bonus is offered, knowing the characteristics of the people who received the bonus and the people who could have received it but did not for whatever reason is relevant. However, personnel data often only record when a benefit was received and not when an individual was eligible for a benefit, which leaves the analyst with the numerator but not the denominator. Expanding the scope of personnel data to more systematically include eligibility will be crucial for effectively examining the impact of numerous personnel policies.

B. Developing Open, Reusable Tools

Recommendation 7: *Improve the understanding of what kinds of questions can be answered by different algorithms and models.*

Recommendation 8: *Invest in developing and sustaining reusable toolkits of algorithms and data, especially those that can be applied across military services and agencies.*

Recommendation 9: *Encourage an open-source, collaborative culture among creators and users of analytic tools throughout the DOD enterprise.*

In an effort to become smart buyers of machine learning analyses, DOD leaders must understand what kinds of questions can be answered by different algorithms and models. With the proliferation of off-the-shelf machine learning software, the capability of a given algorithm can often be overstated or misrepresented. A tool designed to answer a *what* question is likely not suitable to answer a *why* question.

However, within each class of *what*, *how* (or *why*), and *what if* questions, some tools can be used repeatedly and with relatively little adaptation to address a variety of problems. DOD should invest in developing such core, reusable toolkits of algorithms and data. One such example is the FIFE[89] for time-to-event forecasting, together with its application, the RPM (see Section 2.B). Similar toolkits that incorporate the latest causal techniques for assessing retrospective and prospective policy analysis should also be developed.

DOD should especially invest in developing core tools that can be applied to perennial force management issues across the military services. For instance, all the military services face the ongoing challenge of how to set their SRBs. Developing tools that can assist in more systematically identifying the ROI from bonuses would provide immense value to DOD.

To become operational (rather than isolated, proof-of-concept research prototypes), these tools require sustained investment. Code and algorithms should be available to inspect, build upon, and develop. Subject matter experts and stakeholders should be able to validate the veracity and applicability of the results. The algorithms and data feeds also need to be maintained and updated to ensure that results are always current.

[89] The Finite-Interval Forecasting Engine, or FIFE, was developed by IDA at the request of OUSD(P&R) in a manner promoting open-source reuse and replicability.

To facilitate an open environment, where algorithms can be scrutinized, adapted to new defense settings, and modularly combined like building blocks in developing new capabilities, DOD should encourage an open-source culture for these tools throughout the DOD enterprise.

C. Auditing algorithms for Legal, Moral, and Ethical Implications

Recommendation 10: *Adopt best practices for auditing the results of processes using operationalized machine learning or artificial intelligence tools. Actively investigate the legal, moral, and ethical implications of using the results of such algorithms.*

Machine learning algorithms identify patterns within data—but not from a legal or ethical standpoint whether those patterns should be in the data in the first place. If biases exist in promotion decisions, for example, a machine learning algorithm for predicting the service members who are likely be promoted will perpetuate the existing biases. Developing algorithms that are not susceptible to the perpetuation of existing biases is a challenge. Simply omitting information on protected classes, such as gender, race, or religion, does not remove biases, because individuals of a particular class may behave differently in a way that implicitly defines their membership in a protected class through other variables. DOD needs to ensure that specific use cases of machine learning algorithms are carefully audited so that legally protected classes and others are not mistakenly disadvantaged. Best practices for auditing algorithms for bias are steadily maturing and evolving.[90] In operationalizing machine-learning-based algorithms, DOD should actively investigate the legal, moral, and ethical contours for how the algorithm results can be used and judiciously determine appropriate uses.

[90] For instance, the Brookings Institution issued a recent report on best practices for detecting algorithmic biases, and the Institute of Electrical and Electronics Engineers (IEEE) issued a broader report on ethics in machine learning and artificial intelligence. See Nicol Turner Lee, Paul Resnick, and Genie Barton, *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harm* (Washington, DC: Brookings Institution, Center for Technology Innovation, May 22, 2019), https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/; The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*, 1st ed. (IEEE, 2019), https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

This page is intentionally blank.

Appendix A.
Review of Past Studies on Military Career Decisions

While the chapters in the main body of this paper focused on machine learning methodologies that could provide insights for managing the defense personnel workforce, this appendix takes a case-study approach for providing context on a few specific aspects of military retention using more traditional methods. Here, we survey a selection of previous studies, covering the following topics:

1. Natural experiments that reveal the effects of changes in military compensation and in strategies to ensure the validity of natural experiments;
2. Difference-in-differences approaches to retrospective policy evaluation; and
3. Modeling the nature of military exits: voluntary vs. involuntary separation and inter-contract vs. intra-contract separation.

The first two topics fall within the scope of *how* (or *why*) questions for retrospective policy analysis. The third topic emphasizes an aspect of retention modeling that may be relevant to all three classes of questions: the *what*, *how* (or *why*), and *what if*. Contract renewal represents a two-party agreement with the military service and the service member jointly deciding to further the contract relationship. When a contract is not renewed, knowing whether the military service declined to offer a contract or whether the service member declined to accept it is informative. The nature of the separation is also informative: whether the service member separated before the end of a contract period or at the expected time of the contracts end or whether the service member was dismissed for some violation or exited of his or her own volition. If service members have a propensity for one type of exit, do they also have a propensity for other types of exit? If so, how are those propensities correlated? Richer models of retention will involve multiple types of exit and multiple exit paths.

Although these three case studies do not incorporate machine learning, they illustrate a military context for causal analysis into which machine learning methodologies could potentially be applied. We briefly summarize each case study below before covering them in detail.

First, the combat zone tax exemption (CZTE) shields much of the pay earned in a designated combat zone from the federal income tax and from the income tax of most states. The CZTE provides a natural experiment for measuring the effect of increasing after-tax compensation on the probability of reenlistment. However, the experiment is confounded by the presence of stop-loss, the policy that enables the military services to involuntarily keep soldiers in their deployed units even beyond their contracted term of service. One method to estimate the causal effect of the CZTE on reenlistment is to remove the confounding effect of stop-loss on reenlistment. Simon and Warner provide estimates of stop-loss effects that can be adapted for that purpose.[1]

Second, a RAND team led by Jennie Wenger performed a retrospective policy analysis of the Post-9/11 GI Bill.[2] Previous versions of the GI Bill were controversial because while perhaps encouraging people to join the military, the educational benefits were only available after leaving the military. The GI Bill was often criticized as "an incentive to leave." The Post-9/11 GI Bill, which took effect in fiscal year (FY) 2009, attempted to counter that perverse incentive by offering service members the opportunity to transfer some of their educational benefits to their children. Wenger et al. used a difference-in-differences approach to estimate the retention effects of the GI Bill for members with and without dependents. The difference between those two retention effects indicates the extent to which transferability has been successful in muting the incentive to leave the service.

Third, in modeling retention, the nature of separation from the military can be an important but often overlooked factor. Jaffry, Ghulam, and Apostolakis et al. compare voluntary separations ("quits") and involuntary separations ("firings") of sailors from the British Royal Navy with the use of a competing hazards model.[3] Parsing voluntary from involuntary separations requires confidence in how separations are coded in the data. Unfortunately, some ambiguity arises when similarly parsing U.S. military separation codes. Another aspect of modeling retention is distinguishing between separation decisions that occur while a service member is still bound by a contract (intra-contract *attrition*) and separation decisions that occur near the end of the contract (*retention* or *reenlistment*).[4] Demographic or service characteristics related to one type of exit may not apply to another.

[1] Curtis J. Simon and John T. Warner, "Army Re-enlistment During OIF/OEF: Bonuses, Deployment, and Stop-Loss," *Defence and Peace Economics* 21, nos. 5–6 (2010): 507–527, https://doi.org/10.1080/10242694.2010.513488.

[2] Jennie W. Wenger et al., *Are Current Military Education Benefits Efficient and Effective for the Services?* RR-1766-OSD (Santa Monica, CA: RAND Corporation, 2017), https://www.rand.org/pubs/research_reports/RR1766.html.

[3] Shabbar Jaffry, Yaseen Ghulam, and Alexandros Apostolakis, "Analysing Quits and Separations from the Royal Navy," *Defence and Peace Economics* 21, no. 3 (2010): 207–228, https://doi.org/10.1080/10242690903568959.

[4] Reenlistment is a particular action that a service member can take when approaching the end of her contract. Other options include extending for shorter periods of time (e.g., to remain in the military until a college semester begins or in anticipation of a larger reenlistment bonus in the coming fiscal year). Some of these distinctions are described in Matthew S. Goldberg, "A Survey of Enlisted Retention: Models and Findings," in *Report of The Ninth Quadrennial Review of Military Compensation*, vol. III, *Creating Differentials in Military Pay: Special and Incentive Pays* (Washington, DC: Office of the Under Secretary of Defense for Personnel and Readiness, March 2002), 65–134, https://militarypay.defense.gov/Portals/3/Documents/Reports/9th_QRMC_Report_Volumes_I_-_V.pdf.

Within a traditional survival analysis framework, Follmann, Goldberg, and May provide an approach for identifying those different behaviors.[5]

Natural Experiment: CZTE and Stop-Loss

Simon and Warner[6] use probit regression to estimate the effects of bonuses, deployment, and stop-loss policies (among other factors) on reenlistment decisions made by active Army soldiers during FYs 2002 through 2006.[7] They separately analyze soldiers in Zone A (two to six years of service), Zone B (seven to ten years of service), and Zone C (eleven to fourteen years of service), although with lesser emphasis on the third category. They restrict their analysis to some twenty-four Army Military Occupational Specialties (MOSs) that, together, comprise about half of all active Army personnel.

The Stop-Loss Policy

Stop-loss was a controversial policy that the Army applied to involuntarily keep soldiers in the Army while their units were deployed. Notably, this policy could keep soldiers beyond their end of term-of-service (ETS) date at which they could ordinarily leave the Army without penalty and become eligible for benefits from the Department of Veterans Affairs (VA). Stop-loss could compel a soldier to serve from 90 days *before* his or her unit deployed overseas until 90 days *after* the unit returns from deployment.

A common presumption was that most soldiers who were compelled to serve beyond their ETS date would leave at the earliest opportunity. However, a confounding factor was that income received while in a designated combat zone was excluded from federal taxation, and many state taxation laws followed suit.[8] Selected reenlistment bonuses (SRBs) are specifically excluded from taxation and fall under the general rubric of the CZTE.[9] Some soldiers under stop-loss may have reenlisted because the value of their SRB was enhanced by the CZTE to a degree that offset any distaste for service due to stop-loss.

[5] Dean A. Follmann, Matthew S. Goldberg, and Laurie May, “Modeling Spikes in Hazard Rates,” CNA Research Contribution (CRC) 572 (Arlington, VA: Center for Naval Analyses, October 1987); Dean A. Follmann, Matthew S. Goldberg, and Laurie May, “Personal Characteristics, Unemployment Insurance, and the Duration of Unemployment,” *Journal of Econometrics* 45, no. 3 (1990): 351–366, https://doi.org/10.1016/0304-4076(90)90004-D.

[6] Simon and Warner, “Army Re-enlistment During OIF/OEF.”

[7] The Simon and Warner paper is an abridged version of Beth J. Asch et al., *Cash Incentives, Military Enlistment, Attrition, and Reenlistment*, MG-950-OSD (Santa Monica, CA: RAND Corporation, 2010), Chapter 7, https://www.rand.org/pubs/monographs/MG950.html.

[8] Saul Pleeter et al., *Risk and Combat Compensation*, IDA Paper P-4747 (Alexandria, VA: Institute for Defense Analyses, August 2011).

[9] The total exclusion is unlimited for enlisted personnel and warrant officers. It is capped for commissioned officers at the maximum senior enlisted pay.

A 2006 report by the Congressional Budget Office (CBO) “estimated that about 90 percent of those soldiers kept in the Army past their contract expiration date would not reenlist when their stop-loss orders were lifted; instead, they would separate at the earliest possible opportunity.”[10] By contrast, Simon and Warner’s[11] findings were not so pessimistic. They developed a complex set of dummy variables to encode whether a soldier was deployed at the time of his or her reenlistment decision, whether the soldier was under stop-loss at that time, and, if under stop-loss, whether the soldier was in the first fiscal year of stop-loss or was carried over into a second fiscal year at the time of decision.

The Effect of Stop-Loss on Deployed Reenlistment Probabilities

Ideally, we would like to estimate the average probabilities of reenlistment for two groups of deployed soldiers: one that is under stop-loss and the another that is not under stop-loss. Unfortunately, it is not possible to recover those precise probabilities from the Simon and Warner[12] article (nor from the longer RAND report[13]). While the Simon and Warner article reports the respective differences from the average reenlistment probability for a soldier who is neither deployed nor under stop-loss, it does not provide the *levels* of the probabilities. The levels can only be reverse-engineered from the (reported) sample mean probability and the (unreported) sample proportions of soldiers in each two- or three-way category.

The Institute for Defense Analyses (IDA) approximated the two probabilities of interest by assuming that the proportions of soldiers at their ETS date who are deployed or under stop-loss are the same as the overall proportions of soldiers who are deployed or under stop-loss (whether at ETS or not). Table A-1 shows IDA’s estimated probabilities under this assumption.

Two sets of estimates are given for Zone A (two to six years of service) and Zone B (seven to ten years of service) to reflect Simon and Warner’s alternative estimating methods, which are designed to bound the true values. By either estimating method, the probability of reenlistment among deployed soldiers in Zone A is about 53% as large if they are under stop-loss and in Zone B is about 69% as large under stop-loss.

[10] Congressional Budget Office, *Recruiting, Retention, and Future Levels of Military Personnel* (Washington, DC: CBO, October 2006), 34 (footnote 93), http://www.cbo.gov/sites/default/files/109th-congress-2005-2006/reports/10-05-recruiting.pdf.

[11] Simon and Warner, “Army Re-enlistment During OIF/OEF.”

[12] Ibid.

[13] Wenger et al., *Are Current Military Education Benefits Efficient*.

Table A-1. Estimated Probabilities of Reenlistment among Active Soldiers, FYs 2002–2006

Zone	Condition	Probabilities: Model 1	Probabilities: Model 2
A	Deployed, under stop-loss	0.353	0.368
	Deployed, not under stop-loss	0.676	0.683
	Ratio (stop-loss ÷ non-stop-loss)	*52.2%*	*53.9%*
B	Deployed, under stop-loss	0.558	0.571
	Deployed, not under stop-loss	0.816	0.818
	Ratio (stop-loss ÷ non-stop-loss)	*68.4%*	*69.8%*

Source: The IDA calculations in this table were based on Curtis J. Simon and John T. Warner, "Army Re-enlistment During OIF/OEF: Bonuses, Deployment, and Stop-Loss," *Defence and Peace Economics* 21, nos. 5–6 (2010): 507–527, https://doi.org/10.1080/10242694.2010.513488.

For Zone A, it is unclear whether it is more reasonable to extrapolate the *absolute* difference in reenlistment probabilities between being under stop-loss or not (estimated as 32.3 percentage points)[14] or the *relative* difference (estimated as 53% as large under stop-loss). The relative magnitudes of the two adjustments are reversed, depending on whether we are considering an environment in which the reenlistment probabilities are lower or higher than the average during Simon and Warner's[15] sample period. Figure A-1 illustrates this phenomenon. We first hypothesize a stronger economy, implying that the reenlistment probability for deployed service members not under stop-loss declines from the sample average of 67.6% to say 50%. A proportional 53% factor would drop the reenlistment probability for those under stop-loss to 26.1% (not quite cutting the probability in half). Alternatively, an absolute adjustment of 32.3 percentage points would drop the probability even further to 17.7%.

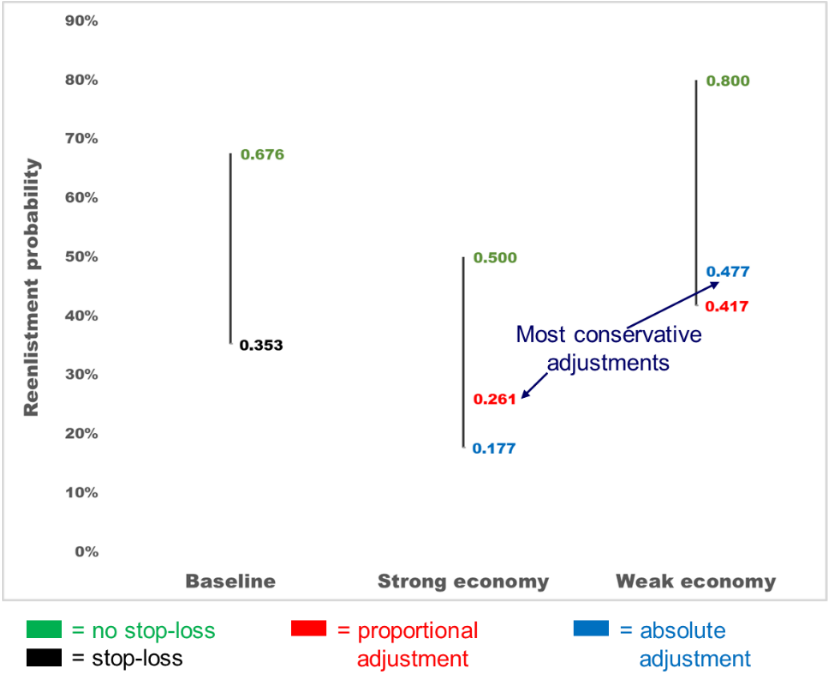
Next, consider the case of a weaker economy so that the reenlistment probability for deployed members not under stop-loss rises from the sample average of 67.6% to say 80%. A proportional 53% factor would drop the reenlistment probability for those under stop-loss to 41.7% (again, not quite cutting the probability in half). In this case, an absolute adjustment of 32.3 percentage points would result in a smaller drop in the probability to 47.7%. The arrows in Figure A-1 indicate a conservative approach that selects the smaller effect of stop-loss in each case (i.e., the proportional adjustment for lower baseline probabilities and the absolute adjustment for higher baseline probabilities).

Source: The IDA calculations in this figure were based on Simon and Warner, "Army Re-enlistment During OIF/OEF."

[14] In Table A-1, this is the Model 1 difference between not under and under stop-loss: 0.676 – 0.353.

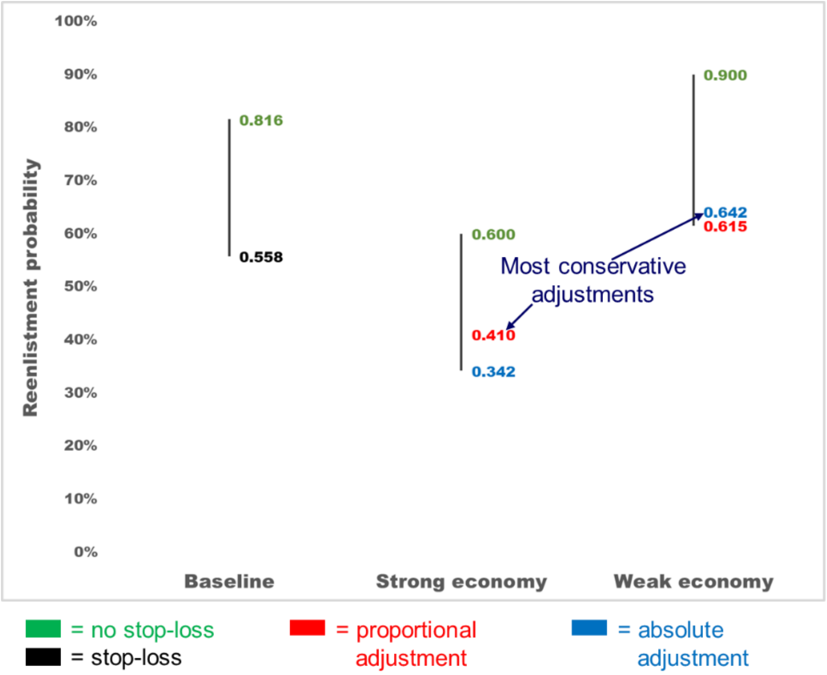
[15] Simon and Warner, "Army Re-enlistment During OIF/OEF."

Figure A-2 illustrates the corresponding analysis for Zone B. Although the precise numbers are different in Zone B (compared to Zone A), the ordering of the effects is identical. The conservative approach involves a 69% proportional factor when the economy is stronger, but a 25.8 percentage-point drop when the economy is weaker.



Source: The IDA calculations in this figure were based on Simon and Warner, "Army Re-enlistment During OIF/OEF."

Figure A-1. Zone A (Years Two to Six)
Adjustments to Reenlistment Probabilities for Stop-Loss



Source: The IDA calculations in this figure were based on Simon and Warner, "Army Re-enlistment During OIF/OEF."

Figure A-2. Zone B (Years Seven to Ten)
Adjustments to Reenlistment Probabilities for Stop-Loss

Adjusting for the Confounding Effect of Stop-Loss

The CZTE provides a natural experiment for measuring the effect of increasing after-tax compensation on the probability of reenlistment. However, the experiment is confounded by the presence of stop-loss because stop-loss generally holds soldiers longer in the combat zone, where they have tax-free earnings. So, stop-loss induces a spurious negative relationship between the CZTE and reenlistment (i.e., when stop-loss is in effect, the value of the CZTE is higher, yet reenlistment rates appear lower).

Figure A-3 illustrates two strategies to adjust for this confounding effect with a directed acyclic graph modeling the causal relationship. Ignore for the moment the availability of an estimate of the effect of stop-loss on reenlistment. If information on stop-loss is unknown, an instrumental variable that is causally unrelated to stop-loss but predicts the value of the CZTE could be used. One such instrumental variable would be the amount of spousal earnings. Spousal earnings help determine a soldier's federal tax bracket and, thus, predicts the value of the CZTE. The dashed portion of Figure A-3 illustrates the instrument variable strategy.[16]

[16] Conversely, we do *not* have to control for the intervening effect of the CZTE when estimating the causal effect of stop-loss on reenlistment rates. For that causal chain, the CZTE is a *mediator* rather than a confounder. See Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (New York, NY: Basic Books, 2018).

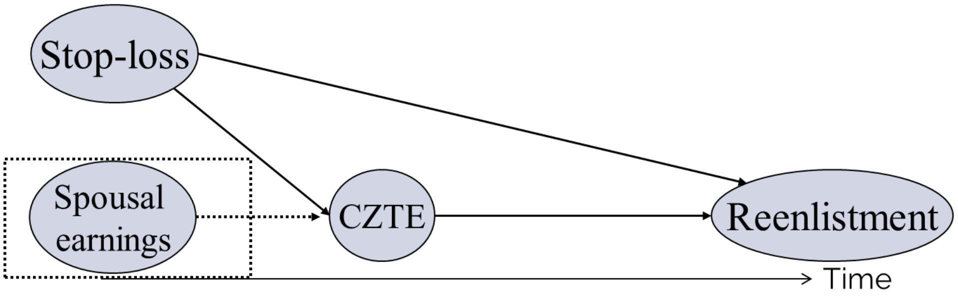


Figure A-3. Causality Graph for Combat Zone Effects on Reenlistment

An alternative strategy first estimates the raw effect of CZTE on reenlistment and then uses auxiliary information to estimate the confounding effect of stop-loss. Specifically, the effects of stop-loss on both CZTE and reenlistment must be removed from the raw effect of CZTE on reenlistment.[17] Such an adjustment is only necessary if stop-loss has a causal effect on both CZTE and on reenlistment as hypothesized. (If there is no causal link between stop-loss and either CZTE or reenlistment, then the adjustment is not necessary.) Since the effect of stop-loss on reenlistment is available from Simon and Warner, it can be combined with the hypothesized signs for the other effects. The resulting adjusted estimate of CZTE on reenlisted will necessarily be larger than the raw estimate since it removes the spurious negative relationship induced by stop-loss.

Retrospective Policy Evaluation: Post-9/11 GI Bill

Wenger et al.[18] formulated and tested several hypotheses regarding the potential impact of the Post-9/11 GI Bill on recruiting and retention.

History of the Post-9/11 GI Bill

The Post-9/11 GI Bill could arguably increase or decrease retention. Earlier versions of the GI Bill encouraged people to join the military. However, since the benefits could only be exercised after leaving the military, they were often criticized as “an incentive to leave.”[19]

This problem was addressed during negotiations in the spring of 2008 between DOD officials and the staffs of the House Armed Services Committee and the Senate Armed

[17] For further details, see for example, the discussion on conducting analysis with omitted variables in a probit regression framework in Jeffrey M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data* (Cambridge, MA: MIT Press, 2002), 470–472. See also William H. Greene, *Econometric Analysis* (New York, NY: Macmillan, 1990), 179.

[18] Wenger et al., *Are Current Military Education Benefits Efficient*, especially Chapter 5.

[19] This sentiment is reported in Bernard Rostker, *I Want You! The Evolution of the All-Volunteer Force*, MG-265-RC (Santa Monica, CA: RAND Corporation, 2006), 512–514, https://www.rand.org/pubs/monographs/MG265.html.

Services Committee. A DOD staffer convinced the committees to add transferability so that service members would not have to separate to exercise their benefits. A report by the CBO recalls the following:

> As the Post-9/11 GI Bill took shape, it developed into the most comprehensive educational benefits package ever offered by the federal government. Concerned that such benefits might motivate service members to leave the military earlier than they would have otherwise, the Department of Defense argued during the drafting of the legislation that the ability to transfer benefits to dependents would be critical to retention goals.[20]

Under transferability, service members can designate some of their benefits for their dependents (combinations of their spouse and one or more children under age twenty-three) if they have completed six years of service and commit to an additional four years.[21] Whereas transferability might offset to some degree the incentive to leave under the GI Bill for members with dependents, no such offset would be available for members without dependents because they have nobody to whom to transfer their benefits.

The legislation began as S. 22, the Post 9/11 Veterans Educational Assistance Act of 2007, introduced by Senator James Webb. A modified version was incorporated into the Supplemental Appropriations Act of 2008, which was enacted as Public Law 110-252 on June 30, 2008. In response to a request from Senator Judd Gregg (ranking member of the Senate Committee on the Budget) in May 2008, the CBO published a cost estimate for S. 22 as it stood on April 23, 2008. Even at the latter date (just over two months before final passage), the bill did not yet contain provisions for transferability. According to CBO's estimate, notwithstanding the direct costs to the VA to pay the benefit, the bill would have increased budgetary costs to DOD by $1.1 billion over the five-year period 2009–2013. That estimate is the difference between the expected cost of increasing reenlistment bonuses to offset the negative retention effects ($6.7 billion), and the estimated savings in enlistment bonuses and other recruiting costs ($5.6 billion).[22]

[20] Congressional Budget Office, "The Post-9/11 GI Bill: Benefits, Choices, and Cost" (Washington, DC: CBO, May 2019), 15, https://www.cbo.gov/system/files/2019-05/55179-Post911GIBill.pdf.

[21] The eligibility rules are complex, and the various DOD and VA websites are difficult to rationalize. An excellent description of program eligibility and benefits is found in Congressional Budget Office, "The Post-9/11 GI Bill: Benefits, Choices, and Cost," 4–5. For example, a member can designate his or her children only after the member has completed ten years of service. A child may be designated only if he or she is under twenty-three years old, although the child may exercise the benefit for an additional three years (until age twenty-six) as long as the he or she is eighteen years or older and has a high school diploma. Finally, the child can continue to exercise the benefit even if no longer a dependent.

[22] Congressional Budget Office, "S.22, the Post 9/11 Veterans Educational Assistance Act of 2008" (Washington, DC: CBO, May 8, 2008), 2, https://www.cbo.gov/sites/default/files/110th-congress-2007-2008/costestimate/s221.pdf. Public Law 110-252 was amended in January 2011 by Public Law 111-377 and again in August 2017 by Public Law 115-48. See Supplemental Appropriations Act, 2008, Pub. L. 110-252, 122 Stat. 2323, 110th Congress (2008), https://www.govinfo.gov/content/pkg/PLAW-110publ252/pdf/PLAW-110publ252.pdf; Post-9/11 Veterans Educational Assistance Improvements Act

Retention Effect Estimates with Difference-in-Differences

Wenger et al.[23] applied a retrospective difference-in-differences approach to separately estimate the retention effects of the GI Bill for members with and without dependents. In turn, the difference between those two retention effects measures the extent to which transferability ameliorates the incentive for members to separate from the military to use their GI Bill benefits.

Specifically, Wenger et al. assembled a sample of members who had between two and a half and four years of completed service, and tracked their retention to five, six, or seven or more years of service. Depending on the retention horizon, service, and component, they estimated that the passage of the Post-9/11 GI Bill depressed year-to-year continuation rates by between one to three percentage points relative to the trend. The effects were slightly smaller in the reserve components.

Next, Wenger et al. introduced in their regression equations an interaction term between the GI Bill effect and an indicator for whether the member had dependents. They found that the depression in continuation rates was only about half as large for members with dependents. The interaction effect—the amelioration of the adverse effect of the GI bill—was largest for the Army.[24]

Modeling the Nature of Military Exits

In a pair of papers published in *Defence and Peace Economics*, Jaffry, Ghulam, and Apostolakis estimated discrete-time hazard models to explain separations of sailors from the Royal Navy.[25] Their earlier paper is among the few that distinguish between voluntary separations ("quits") and involuntary separations ("firings") from military service.

Most enlisted sailors (called "ratings" in the Royal Navy) sign up between ages sixteen and eighteen. They are assigned to one of five branches: operations and warfare; engineering; supply, logistics, and hospitality; fleet air; or medical. They generally remain in their assigned branch for their entire careers.

of 2010, Pub. L. 111-377, 124 Stat. 4106, 111th Congress (2011), https://www.govinfo.gov/content/pkg/PLAW-111publ377/pdf/PLAW-111publ377.pdf; Harry W. Colmery Veterans Educational Assistance Act of 2017, Pub. L. 115-48, 131 Stat. 973, 115th Congress (2017), https://www.congress.gov/115/plaws/publ48/PLAW-115publ48.pdf.

[23] Wenger et al., *Are Current Military Education Benefits Efficient*.

[24] Ibid. See especially (1) Table 5.7 and the surrounding discussion on pages 50–51 and (2) Table C.4, page 99 in Appendix C.

[25] Jaffry, Ghulam, and Apostolakis, "Analysing Quits and Separations from the Royal Navy"; Shabbar Jaffry, Yaseen Ghulam, and Alexandros Apostolakis, "Explaining Early Exit Rates from the Royal Navy," *Defence and Peace Economics* 24, no. 4 (2013): 339–369, https://doi.org/10.1080/10242694.2012.695035.

Upon joining the Royal Navy, sailors sign a "full-service contract" that lasts twenty-two years. Notwithstanding that contract length, sailors may voluntarily separate two and a half years after the end of initial training, provided they give twelve months advance notice. Jaffry, Ghulam, and Apostolakis refer to these planned, *voluntary* separations as "quits." They also recognize and in their 2010 paper separately model *involuntary* separations (i.e., "exits through compassionate, medical, dismissals, unsuitability, and other reasons"[26]).

Just as the authors hypothesized, separation rates dropped precipitously for individuals with nineteen to twenty-three years of service, the period of time immediately preceding eligibility for retirement.[27] That finding is analogous to the well-documented drop in attrition from all branches of the U.S. military during the five or so years leading up to retirement eligibility after twenty years of service.

Competing Risk Models

Jaffry, Ghulam, and Apostolakis[28] estimate a competing risk model between the different forms of separation. This framework assumes the following: each sailor has a date when he or she will voluntarily separate, and each sailor has another date (almost certainly different) when he or she would separate involuntarily. In the data, however, we observe, at most, the earlier of those two dates or possibly neither date if the sailor is still serving at the end of the data period (i.e., the separation data are *censored*).

The authors model each type of risk in a discrete-time hazard framework, where the probability of either type of separation occurring in a year is conditional on having survived to the beginning of that year (also known as the *hazard*) and is expressed as an extreme-value function of a vector of covariates.[29]

[26] Jaffry, Ghulam, and Apostolakis, "Analysing Quits and Separations from the Royal Navy," 208.

[27] Although the exposition in Jaffry, Ghulam, and Apostolakis is unclear, retirement eligibility from the Royal Navy appears to have accrued at around twenty-four years of service during the study period (April 1996 through June 2002) (see Jaffry, Ghulam, and Apostolakis, "Analysing Quits and Separations from the Royal Navy," 220). Armed forces pensions in the United Kingdom have since been revised. The current system, Armed Forces Pensions Scheme (AFPS15), took effect in April 2015 (see "Guidance for Armed Forces Pensions," Government of the United Kingdom, 12 December 2012, last updated March 4, 2020, https://www.gov.uk/guidance/pensions-and-compensation-for-veterans).

[28] Jaffry, Ghulam, and Apostolakis, "Analysing Quits and Separations from the Royal Navy."

[29] Methodologically, the authors use the complementary log-log transformation rather than the more familiar logistic or probit transformations. The complementary log-log transformation has the mathematical property of being the unique way to formulate discrete realization of an underlying continuous proportional hazard model. See John D. Kalbfleisch and Ross L. Prentice, *The Statistical Analysis of Failure Time Data* (New York, NY: John Wiley & Sons, Inc. 1980), 37, 99. The result originally goes back to R. L. Prentice and L. A. Gloeckler, "Regression Analysis of Grouped Survival Data with Applications to Breast Cancer Data," *Biometrics* 34 (March 1978): 57–67, https://www.jstor.org/stable/pdf/2529588.pdf.

The analysis highlights several patterns that are also present in the U.S. military. For instance, separation rates are negatively associated with the civilian unemployment rate. Higher unemployment makes remaining in the Royal Navy an attractive option. Conversely, separation rates are positively associated with changes in an index of wages in the United Kingdom. The authors also report that male sailors are more likely than female sailors to quit or separate due to lack of promotion. Among male sailors, married ones are less likely to quit. Among female sailors, the opposite is true. Also, the propensity to quit is higher with demanding operational tempos.

While the authors express no difficulty in distinguishing voluntary from involuntary separations in their data, several Interservice Separation Codes (ISCs) for the U.S. military are not as easily categorized. Some can be unambiguously categorized as a voluntary separation (ISC 1001: Expiration of term of service) or as an involuntary separation (ISC 1073: Court martial). For others, the "fired" vs. "quit" distinction is not clear from the separation code. For example, when a service member is dismissed for "failure to meet weight or body fat standards" (ISC 1017), it is not clear whether the service member desired to remain in the military but was dismissed (involuntary separation) or whether the service member perhaps gave up on weight control and neglected an exercise routine as an intentional way to exit the military with minimal consequences (voluntary separation). Resolving this ambiguity to apply a competing risk model to the U.S. military would require labeling each exit as either voluntary or involuntary.

Inter- and Intra-Contract Exit: Spikes in Hazard Rates

In a pair of papers, Follmann, Goldberg, and May developed an approach that allows an individual's hazard of exiting to have "spike event."[30] The first paper (in 1987) modeled the attrition of Marine Corps reservists. The spike event was the end of a Marine's initial enlistment contract, which generally occurred at the end of the fourth year (though a smaller number of contracts were for six years). Identifying the contract end date is important because the survival probability drops precipitously at the end of a Marine's contract. The second paper (in 1990) modeled the duration of unemployment in the civilian labor force, where a similar spike occurs at the expiration of unemployment insurance benefits.[31]

[30] Follmann, Goldberg, and May, "Modeling Spikes in Hazard Rates"; Follmann, Goldberg, and May, "Personal Characteristics, Unemployment Insurance." A related approach is found in John T. Warner and Gary Solon, "First-Term Attrition and Reenlistment in the U.S. Army," in *Military Compensation and Personnel Retention: Models and Evidence*, ed. Curtis L. Gilroy, David K. Horne, and D. Alton Smith (Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, 1991), 243–281, https://apps.dtic.mil/dtic/tr/fulltext/u2/a363401.pdf.

[31] Methodologically, the later paper incorporates the complementary log-log hazard function to model behavior at the spike event (the earlier paper uses a logistic function).

The authors first estimate a Weibull survival model for a cohort of Marine Corps reservists. Although the Weibull model seemed to fit the data reasonably well, it did not account for a spike in the attrition rate at between forty-eight and fifty months of service. Most of the enlistment contracts were for a period of four years or precisely forty-eight months. However, unlike active service members, reservists may not formally announce their departure at the conclusion of the contract. The command may presume that the Marine is intending to continue service until a few monthly drill weekends are missed and the command eventually concludes that he has dropped out. Thus, the administrative data may continue to include a Marine reservist for a couple of extra months without officially acknowledging attrition. After fifty months, attrition patterns return to a smooth, gradual pattern, just as before the spike.

To accommodate the two attrition patterns, Follmann, Goldberg, and May[32] developed a likelihood function that decomposes attrition into a three-month "spike" period (forty-eight to fifty months), and a Weibull regression (see Table A-2) for the remainder of the time scale before and after the spike period.[33] Modeling inter-contract exit (the "spike") separately from intra-contract exit reveals some interesting patterns. Specification (1) in Table A-2 shows the Weibull regression results when the entire period is modeled together. Specifications (2) and (3) show the Weibull regression results during "non-spike" periods and the logit regression at the "spike."

Table A-2. Weibull-Logit Model of Attrition among Marine Corps Reservists

Covariate	(1) All periods (Weibull regression)	(2) Non-spike periods (Partial Weibull regression)	(3) Spike period (Logit regression)
Age	-0.018*	-0.017*	-0.070
Married	0.226*	0.236*	0.010
High school degree	-0.179*	-0.194*	0.011
White	-0.041	-0.054	0.407*
Male	-0.015	-0.057	1.544*
Contract end year four			1.548*
Weibull shape parameter	1.329*	1.298*	

Source: Dean A. Follmann, Matthew S. Goldberg, and Laurie May, "Modeling Spikes in Hazard Rates," CNA Research Contribution (CRC) 572 (Arlington, VA: Center for Naval Analyses, October 1987).

Note: Based on data through six years of service. The spike occurs at the end of four years of service.

* Statistically significant at the 1% level.

[32] Follmann, Goldberg, and May, "Modeling Spikes in Hazard Rates."

[33] Covariates enter multiplicatively into the Weibull regression, so that the hazards at any two points in time remain in the same proportion for two individuals with unchanging, but different, covariate values.

In addition, race and gender significantly affect the probability of leaving during the spike period but do not significantly affect the probability of leaving outside the spike period. These results imply that the hazard functions are *not* proportional through time for Marines (i.e., the commonly invoked proportional hazards assumption is violated). Further, the coefficient estimates of the two Weibull models—the standard model for all of the months of data and the partial model that excludes months forty-eight through fifty—are nearly identical. The behavior *away from* the spike periods dominates the standard model because those observations are so much more numerous than the observations *during* the spike periods. For example, the positive effects of race and gender on the probability of separating at the end of a four-year contract are lost when pooling the spike and non-spike periods in the standard model.

Appendix B.
Illustrations

Figures

Tables

This page is intentionally blank.

Appendix C.
References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program.” *Journal of the American Statistical Association* 105, no. 490 (2010): 493–505. https://doi.org/10.1198/jasa.2009.ap08746.

ACM A.M. Turing Award. “Judea Pearl.” http://amturing.acm.org/award_winners/pearl_2658896.cfm.

Angrist, Joshua D. “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records.” *The American Economic Review* 80, no. 3 (June 1990): 313–336. https://www.jstor.org/stable/2006669.

Arcidiacono, Peter, and Paul B. Ellickson. “Practical Methods for Estimation of Dynamic Discrete Choice Models.” *Annual Review of Economics.* 3, no. 1 (2011): 363–394. https://www.annualreviews.org/doi/abs/10.1146/annurev-economics-111809-125038.

Asch, Beth J., Michael G. Mattock, and James Hosek. *A New Tool for Assessing Workforce Management Policies over Time: Extending the Dynamic Retention Model*. RR-113-OSD. Santa Monica, CA: RAND Corporation, 2013. https://www.rand.org/pubs/research_reports/RR113.html.

Asch, Beth J., Paul Heaton, James Hosek, Francisco Martorell, Curtis Simon, and John T. Warner. *Cash Incentives, Military Enlistment, Attrition, and Reenlistment*. MG-950-OSD. Santa Monica, CA: RAND Corporation, 2010. https://www.rand.org/pubs/monographs/MG950.html.

Athey, Susan, and Guido W. Imbens. “Machine Learning Methods Economists Should Know About.” *Annual Review of Economics* 11 (August 2019): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.

Athey, Susan, and Guido W. Imbens. “The State of Applied Econometrics: Causality and Policy Evaluation.” *Journal of Economic Perspectives* 31, no. 2 (Spring 2017): 3–32. https://www.aeaweb.org/articles?id=10.1257/jep.31.2.3.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. *Matrix Completion Methods for Causal Panel Data Models*. NBER Working Paper No. 25132. Cambridge, MA: National Bureau of Economic Research, October 2018. https://www.nber.org/papers/w25132.pdf.

Austin, Peter C. “An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies.” *Multivariate Behavioral Research* 46, no. 3 (2011): 399–424. doi:10.1080/00273171.2011.568786.

Bellman, Richard. *Eye of the Hurricane: An Autobiography*. Hackensack, NJ: World Scientific Publishing Co., Inc., 1984.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *The Review of Economic Studies* 81, no. 2 (April 2014): 608–650. https://www.jstor.org/stable/43551575.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall/CRC, 1984.

Breiman, Leo. "Random Forests." *Machine Learning* 45, no. 1 (2001): 5–32. https://link.springer.com/article/10.1023/A:1010933404324.

Card, David. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review* 43, no. 2 (1990): 245–257. https://doi.org/10.1177/001979399004300205.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21, no. 1 (February 2018): C1–C68. https://doi.org/10.1111/ectj.12097.

Chollet, François. *Deep Learning with Python*, Shelter Island, NY: Manning Publications 2018.

Congressional Budget Office. "S.22, the Post 9/11 Veterans Educational Assistance Act of 2008." Washington, DC: CBO, May 8, 2008. https://www.cbo.gov/sites/default/files/110th-congress-2007-2008/costestimate/s221.pdf.

Congressional Budget Office. "The Post-9/11 GI Bill: Benefits, Choices, and Cost." Washington, DC: CBO, May 2019. https://www.cbo.gov/system/files/2019-05/55179-Post911GIBill.pdf.

Congressional Budget Office. *Recruiting, Retention, and Future Levels of Military Personnel*. Washington, DC: CBO, October 2006. http://www.cbo.gov/sites/default/files/109th-congress-2005-2006/reports/10-05-recruiting.pdf.

Cummins, Nicholas, Julien Epps, and Eliathamby Ambikairajah. "Spectro-Temporal Analysis of Speech Affected by Depression and Psychomotor Retardation." In *2013 IEEE International Conference on Acoustics, Speech, and Signal Processing: Proceedings*, 7542–7546. IEEE, 2013. https://ieeexplore.ieee.org/document/6639129.

Department of Defense. *Summary of the 2018 National Defense Strategy of the United States of America: Sharpening the American Military's Competitive Edge*. Washington, DC: Office of the Secretary of Defense, 2018. https://dod.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf.

Doyle, Colin M. *The Military Career Analysis Model: Theory, Methodology, Estimation, and Application of a Unified Model of Military Personnel Accession, Retention, motion, and Life-Cycle Cost.* IDA Document P-5264. Alexandria, VA: Institute for Defense Analyses, June 2015.

Ermon, Stefano, Yexiang Xue, Russell Toth, Bistra Dilkina, Richard Bernstein, Theodoros Damoulas, Patrick Clark, et al. "Learning Large-Scale Dynamic Discrete Choice Models of Spatio-Temporal Preferences with Application to Migratory Pastoralism in East Africa." In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 644–650. Palo Alto, CA: AAAI Press, 2015. https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9725/9308.

Follmann, Dean A., Matthew S. Goldberg, and Laurie May. "Personal Characteristics, Unemployment Insurance, and the Duration of Unemployment." *Journal of Econometrics* 45, no. 3 (1990): 351–366. https://doi.org/10.1016/0304-4076(90)90004-D.

Follmann, Dean A., Matthew S. Goldberg, and Laurie May. "Modeling Spikes in Hazard Rates." CNA Research Contribution (CRC) 572. Arlington, VA: Center for Naval Analyses, October 1987.

Friedman, Jerome H. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29, no. 5 (October 2001): 1189–1232. https://www.jstor.org/stable/2699986.

Frisch, Ragnar. "Editorial." *Econometrica* 1, no. 1 (January 1933): 1–4. https://www.jstor.org/stable/1912224.

Gal, Yarin, and Zoubin Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." In *Proceedings of the 33rd International Conference on Machine* Learning – Volume 48, edited by Maria Florina Balcan and Kilian Q. Weinberger, 1050–1059. JMLR.org, 2016. http://proceedings.mlr.press/v48/gal16.pdf.

Goldberg, Matthew S. "A Survey of Enlisted Retention: Models and Findings." In *Report of The Ninth Quadrennial Review of Military Compensation*. Vol. III, *Creating Differentials in Military Pay: Special and Incentive Pays*, 65–134. Washington, DC: Office of the Under Secretary of Defense for Personnel and Readiness, March 2002. https://militarypay.defense.gov/Portals/3/Documents/Reports/9th_QRMC_Report_Volumes_I_-_V.pdf.

Gotz, Glenn A., and John J. McCall. *A Dynamic Retention Model for Air Force Officers: Theory and Estimates.* R-3028-AF. Santa Monica, CA: RAND Corporation, December 1984. https://www.rand.org/pubs/reports/R3028.html.

Gotz, Glenn A., and John J. McCall. *A Sequential Analysis of the Air Force Officer's Retirement Decision*. N-1013-1-AF. Santa Monica, CA: RAND Corporation, 1979. https://www.rand.org/pubs/notes/N1013-1.html.

Gotz, Glenn A., and John J. McCall. *Estimating Military Personnel Retention Rates: Theory and Statistical Method.* R-2541-AF. Santa Monica, CA: RAND Corporation, June 1980. https://www.rand.org/pubs/reports/R2541.html.

Government of the United Kingdom, "Guidance for Armed Forces Pensions," 12 December 2012. Last updated March 4, 2020. https://www.gov.uk/guidance/pensions-and-compensation-for-veterans.

Greene, William H. *Econometric Analysis*. New York, NY: Macmillan, 1990.

Harry W. Colmery Veterans Educational Assistance Act of 2017. Pub. L. 115-48. 131 Stat. 973. 115th Congress (2017). https://www.congress.gov/115/plaws/publ48/PLAW-115publ48.pdf.

Hartford, Jason, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. "Deep IV: A Flexible Approach for Counterfactual Prediction." In *Proceedings of the 34th International Conference on Machine Learning* – Volume 70, edited by Doina Precup and Yee Whye Teh, 1414–1423. JMLR.org, 2017. http://proceedings.mlr.press/v70/hartford17a.html.

Heckman, James J. "The Scientific Model of Causality." *Sociological Methodology* 35, no. 1 (August 2005): 1–97. https://doi.org/10.1111/j.0081-1750.2006.00164.x.

Heckman, James J., and Robert J. Willis. "Estimation of a Stochastic Model of Reproduction: An Econometric Approach." In *Household Production and Consumption*, edited by Nestor E. Terleckyj, 99–146. Cambridge, MA: National Bureau of Economic Research (NBER), 1976. https://www.nber.org/books/terl76-1.

Heskes, Tom. "Practical Confidence and Prediction Intervals." In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, edited by Michael I. Jordan, Yann LeCun, and Sara A. Solla, 176–182. Cambridge, MA: The MIT Press, December 1996. http://papers.nips.cc/paper/1306-practical-confidence-and-prediction-intervals.pdf.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71, no. 4 (July 2003): 1161–1189. https://doi.org/10.1111/1468-0262.00442.

Horwitz-Martin, Rachelle L., Thomas F. Quatieri, Elizabeth Godoy, and James R. Williamson. "A Vocal Modulation Model with Application to Predicting Depression Severity." In *Proceedings of the 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 247–253. Stoughton, WI: The Printing House, 2016. https://ieeexplore.ieee.org/document/7516268.

Hotz, V. Joseph, and Robert A. Miller. "Conditional Choice Probabilities and the Estimation of Dynamic Models." *The Review of Economic Studies* 60, no. 3 (1993): 497–529. https://www.jstor.org/stable/2298122.

Huff, Jared, Mikhail Smirnov, Greggory Schell, and James Grefer. *Estimating the Retention Effects of Continuation Pay*. Arlington, VA: Center for Naval Analyses (CNA), April 2018. https://www.cna.org/CNA_files/PDF/DRM-2018-U-017177-Final.pdf.

Imbens, Guido W. "Instrumental Variables: An Econometrician's Perspective." *Statistical Science* 29, no. 3 (August 2014): 323–358. https://www.jstor.org/stable/43288511.

Jaffry, Shabbar, Yaseen Ghulam, and Alexandros Apostolakis. "Explaining Early Exit Rates from the Royal Navy." *Defence and Peace Economics* 24, no. 4 (2013): 339–369. https://doi.org/10.1080/10242694.2012.695035.

Jaffry, Shabbar, Yaseen Ghulam, and Alexandros Apostolakis. "Analysing Quits and Separations from the Royal Navy." *Defence and Peace Economics* 21, no. 3 (2010): 207–228. https://doi.org/10.1080/10242690903568959.

Jaggi, Martin. “An Equivalence between the Lasso and Support Vector Machines.” In *Regularization, Optimization, Kernels, and Support Vector Machines*, edited by Johan A. K. Suykens, Marco Signoretto, and Andreas Argyriou, 1–26. Boca Raton, FL: CRC Press, 2015.

Kalbfleisch, John D., and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. New York, NY: John Wiley & Sons, Inc. 1980.

Kleiner, Ariel, Ameet Talwalkar, Purnamrita Sarkar, and Michael I. Jordan. “A Scalable Bootstrap for Massive Data.” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 76, no. 4 (September 2014): 795–816. https://www.jstor.org/stable/24774569.

Lee, Nicol Turner, Paul Resnick, and Genie Barton. *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harm*. Washington, DC: Brookings Institution, Center for Technology Innovation, May 22, 2019. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

Mazumder, Rahul, Trevor Hastie, and Robert Tibshirani. “Spectral Regularization Algorithms for Learning Large Incomplete Matrices.” *Journal of Machine Learning Research* 11 (2010): 2287–2322. https://dl.acm.org/doi/10.5555/1756006.1859931.

McGinnis, Ellen W., Steven P. Anderau, Jessica Hruschak, Reed D. Gurchiek, Nestor L. Lopez-Duran, Kate Fitzgerald, Katherine L. Rosenblum, et al. “Giving Voice to Vulnerable Children: Machine Learning Analysis of Speech Detects Anxiety and Depression in Early Childhood.” *IEEE Journal of Biomedical and Health Informatics* 23, no. 6 (November 2019): 2294–2301. https://ieeexplore.ieee.org/document/8700173.

Microsoft. “Failure Modes in Machine Learning.” November 2019. https://docs.microsoft.com/en-us/security/failure-modes-in-machine-learning.

Ng, Andrew Y., and Stuart J. Russell. “Algorithms for Inverse Reinforcement Learning.” In *Proceedings of the Seventeenth International Conference on Machine Learning*, edited by Pat Langley, 663–670. San Francisco, CA: Morgan Kaufmann Publishers, Inc. 2000. https://ai.stanford.edu/~ang/papers/icml00-irl.pdf.

Norets, Andriy. “Estimation of Dynamic Discrete Choice Models Using Artificial Neural Network Approximations.” *Econometric Reviews* 31, no. 1 (2012): 84–106. https://doi.org/10.1080/07474938.2011.607089.

Pearce, Tim, Felix Liebfried, Alexandra Brintrup, Mohamed Zaki, and Andy Neel. “Uncertainty in Neural Networks: Approximately Bayesian Ensembling.” arXiv:1810.05546v5. 2019.

Pearce, Tim, Mohamed Zaki, Alexandra Brintrup, and Andy Neely. “High-Quality Prediction Intervals for Deep Learning: A Distribution-Free, Ensembled Approach.” In *Proceedings of the 35th International Conference on Machine Learning* – Volume 80, edited by Jennifer Dy and Andreas Krause, 4075–4084. Red Hook, NY: Curran Associates, Inc., 2018. http://proceedings.mlr.press/v80/pearce18a/pearce18a.pdf.

Pearl, Judea, and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books, 2018.

Pechacek, Julie, Alan Gelder, Amrit Romana, Ethan Novak, Kathy Conley, Cheryl Green, Dina Eliezer, et al. *Considerations for Implementing a Defense Personnel Research Environment.* IDA Document P-9254. Alexandria, VA: Institute for Defense Analyses, September 2018.

Pechacek, Julie, Alan Gelder, Ethan Novak, Amrit Romana, Paul Richanbach, Kathy Conley, George Kennedy, and Cheryl Green. *User Requirements for the Enterprise Data to Decisions Information Environment.* IDA Document NS D-9139. Alexandria, VA: Institute for Defense Analyses, August 2018.

Pleeter, Saul, Alexander O. Gallo, Brandon R. Gould, Maggie X. Li, Shirley H. Liu, Curtis J. Simon, Carl F. Witschonke, and Stanley A. Horowitz. *Risk and Combat Compensation*. IDA Paper P-4747. Alexandria, VA: Institute for Defense Analyses, August 2011.

Polich, J. Michael, James N. Dertouzos, and S. James Press. *The Enlistment Bonus Experiment*. R-3353-FMP. Santa Monica, CA: RAND Corporation, 1986. https://www.rand.org/pubs/reports/R3353.html.

Post-9/11 Veterans Educational Assistance Improvements Act of 2010. Pub. L. 111-377. 124 Stat. 4106. 111th Congress (2011). https://www.govinfo.gov/content/pkg/PLAW-111publ377/pdf/PLAW-111publ377.pdf.

Prentice, R. L. and L.A. Gloeckler. "Regression Analysis of Grouped Survival Data with Applications to Breast Cancer Data." *Biometrics* 34 (March 1978): 57–67. https://www.jstor.org/stable/pdf/2529588.pdf.

Rosenbaum, Paul R., and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70, no. 1 (01 April 1983): 41–55. https://doi.org/10.1093/biomet/70.1.41.

Rostker, Bernard D. *I Want You! The Evolution of the All-Volunteer Force*. MG-265-RC. Santa Monica, CA: RAND Corporation, 2006. https://www.rand.org/pubs/monographs/MG265.html.

Rust, John. "Numerical Dynamic Programming in Economics." In *Handbook of Computational Economics*, Volume 1, edited by Hans M. Amman, David A. Kendrick, and John Rust, 619–729. Amsterdam, The Netherlands: Elsevier Science B.V., 1996. https://doi.org/10.1016/S1574-0021(96)01016-7.

Rust, John. "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher." *Econometrica* 55, no. 5 (September 1987): 999–1033. https://www.jstor.org/stable/1911259.

Scheidegger, Simon, and Ilias Bilionis. "Machine Learning for High-Dimensional Dynamic Stochastic Economies." *Journal of Computational Science* 33 (April 2019): 68–82. https://doi.org/10.1016/j.jocs.2019.03.004.

Sharma, Mohit, Kris M. Kitani, and Joachim Groeger. "Inverse Reinforcement Learning with Conditional Choice Probabilities." arXiv:1709.07597v1. 2017.

Simon, Curtis J., and John T. Warner. “Army Re-enlistment During OIF/OEF: Bonuses, Deployment, and Stop-Loss.” *Defence and Peace Economics* 21, nos. 5–6 (2010): 507–527. https://doi.org/10.1080/10242694.2010.513488.

Supplemental Appropriations Act, 2008. Pub. L. 110-252. 122 Stat. 2323. 110th Congress (2008). https://www.govinfo.gov/content/pkg/PLAW-110publ252/pdf/PLAW-110publ252.pdf.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*, 1st ed. IEEE, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/ autonomous-systems.html.

U.S. Department of Defense. “Our Story.” Accessed December 9, 2019. https://www.defense.gov/our-story/.

Wager, Stefan, and Susan Athey. “Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests.” *Journal of the American Statistical Association* 113, no. 523 (2018): 1228–1242. https://doi.org/10.1080/01621459.2017.1319839.

Warner, John T., and Gary Solon. “First-Term Attrition and Reenlistment in the U.S. Army.” In *Military Compensation and Personnel Retention: Models and Evidence*, edited by Curtis L. Gilroy, David K. Horne, and D. Alton Smith, 243–281. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, 1991. https://apps.dtic.mil/dtic/tr/fulltext/u2/a363401.pdf.

Wenger, Jennie W., Trey Miller, Matthew D. Baird, Peter Buryk, Lindsay Daugherty, Marlon Graf, Simon Hollands, et al. *Are Current Military Education Benefits Efficient and Effective for the Services?* RR-1766-OSD. Santa Monica, CA: RAND Corporation, 2017. https://www.rand.org/pubs/research_reports/RR1766.html.

Wooldridge, Jeffrey M. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, 2002.

Wright, Philip G. *The Tariff on Animal and Vegetable Oils*. New York, NY: Macmillan Company, 1928.

Zhou, Quan, Wenlin Chen, Shiji Song, Jacob R. Gardner, Kilian Q. Weinberger, and Yixin Chen. “A Reduction of the Elastic Net to Support Vector Machines with an Application to GPU Computing.” In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 3210–3216. Palo Alto, CA: AAAI Press, 2015. http://www.cs.cornell.edu/~kilian/papers/aaai15_sven.pdf.

Ziebart, Brian D., Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. “Maximum Entropy Inverse Reinforcement Learning.” In *Proceedings of the 23rd National Conference on Artificial Intelligence* – Volume 3, edited by Anthony Cohn, 1433–1438. Palo Alto, CA: AAAI Press, 2008, https://www.aaai.org/Papers/AAAI/2008/ AAAI08-227.pdf.

This page is intentionally blank.

Appendix D.
Glossary

Bootstrapping

A method for quantifying uncertainty based on the concept that the available sample is representative of the underlying data from which it is drawn. The researcher repeatedly selects observations from the available sample to identify distributional properties of the larger data that it represents.

Causal Effect

The impact of a given contributing condition on a selected outcome (e.g., “the causal effect of policy *x* on outcome *y* is a ten percent increase in *y*, all else equal”).

Causal Forests

A variant of the random forests method that can identify treatment effects within subgroups while maintaining the conditions needed to produce valid confidence intervals for those treatment effects.

Causal Model

A scientific framework that describes or demonstrates how different environments, policies, processes, or conditions affect specific outcomes.

Confounded

A situation where a direct estimation of a given causal effect is not possible because the influences of multiple factors are combined in the observed or available data.

Difference-in-Differences

A method for estimating the causal effect of a treatment when individuals in a treatment and a control group are observed repeatedly over time—before and after administration of the treatment.

Double Machine Learning

A method that permits the valid estimation of statistical models that have large numbers of covariates relative to the number of observations. The method uses machine learning techniques twice: first to focus the model on the most salient covariates and second to remove the statistical bias introduced into the model in the first step.

Dynamic Programming

A mathematical optimization exercise that identifies an optimal decision path. Commonly used in economics to implement agent-based models in which synthetic individuals chose their actions in a highly stylized, defined context.

External Validity

The extent to which a piece of evidence supports a claim about cause and effect and is generalizable beyond the context of the particular group, time period, or context studied.

Gradient Boosted Trees

A tree-based machine learning method that sequentially builds tree models in a layered manner, with each successive tree using the portion of the data that the previous tree did not explain.

Hazard Models

A family of statistical methods used for estimating time to an event or the chance that an event occurs in a given future period (e.g., death, equipment failure, separation from military service) based on the characteristics of a person or object.

“How” or “Why” Questions

Questions that seek to explain why or how observed outcomes occurred by identifying cause-and-effect relationships from events or conditions in the past. A question seeking to identify causal effects.

Hyperparameter

An aspect of a machine learning algorithm which the researcher specifies (in contrast to *parameters*, which are determined algorithmically).

Instrumental Variables

A method to estimate the causal effect of a treatment when certain types of bias are present in direct estimation, using a variable (known as an *instrument*) that does not directly affect the outcome of interest but does affect the probability that an individual receives the treatment.

Internal Validity

The extent to which a piece of evidence supports a claim about cause and effect, within the context of a particular group, time period, or context.

Inverse Reinforcement Learning

A family of machine learning methods that seek to infer the underlying objective and reward system that motivates observed behavior.

Natural Experiment
(or Quasi Experiment)

Natural or unplanned events that resemble or approximate experimental conditions. For example, the implementation of a new policy or law that applies in one jurisdiction (the treatment group) but not in an otherwise similar jurisdiction (the control group).

Neural Networks
A family of machine learning methods that are built using collections of decision points, called neurons, that can be combined and ordered in various ways (e.g., in layers) to produce an overall classification or assessment.

Overfitting
A type of model generation failure that occurs when an algorithm identifies patterns between the inputs and outcome that exist within the data used to develop the algorithm but that do not exist generally (i.e., patterns that are not externally valid).

Panel Data
Data for which people or objects are observed repeatedly over time.

Propensity Score
The probability that an individual receives a treatment based on the characteristics of the individual that can be observed in the data.

Random Forests
A tree-based machine learning method wherein an algorithm builds numerous model "trees," each using a random subset of the available data and a random subset of the variables describing the data, to make predictions that are less prone to overfitting.

Randomized Control Trial
A method for establishing the causal effect of a treatment whereby individuals or objects are randomly assigned to otherwise identical test and control groups.

Regression Discontinuity
A method for estimating the causal effect of a policy that has a strict eligibility cutoff (perhaps in terms of a person's age, income, or years of service). The method assesses whether the eligibility cutoff coincides with a jump in the outcome of interest.

Synthetic Control
A variant of the difference-in-differences method that compares the treatment group to a weighted combination of several potential control groups constructed to best resemble the treatment group (instead of comparing the treatment group to a single control group or to a set of equally weighted control groups that may not closely match the various characteristics of the treatment group).

Tree-Based Models
A family of machine learning methods that split observations based on their characteristics into successively finer groups that share common outcomes.

Unconfounded
A situation where the direct estimation of a causal effect is possible (perhaps after controlling for known confounding factors).

“What” Questions

Questions that seek to describe—not to explain—past or present conditions or to forecast future conditions using data patterns to generate predictions, groupings, classifications, tabulations and to identify trends and correlations.

“What if” Questions

Hypothetical questions that ask what will happen in the future if important changes are made or if certain events occur.

Appendix E. Abbreviations

AAAI	Association for the Advancement of Artificial Intelligence
ACM	Association for Computing Machinery
AFPS	Armed Forces Pensions Scheme
AIP	Assignment Incentive Pay
BRS	Blended Retirement System
CBO	Congressional Budget Office
CNA	Center for Naval Analyses
CRC	CNA Research Contribution
CSP	Career Satisfaction Program
CZTE	combat zone tax exemption
DERM	Dynamic Econometric Retention Model
DHA	Defense Health Agency
DMDC	Defense Manpower Data Center
DOD	Department of Defense
DRM	Dynamic Retention Model
ETS	end of term-of-service
FFN	feedforward neural network
FIFE	Finite-Interval Forecasting Engine
FY	fiscal year
IDA	Institute for Defense Analyses
IEEE	Institute of Electrical and Electronics Engineers
ISC	Interservice Separation Code
JAMRS	Joint Advertising, Marketing, Research, and Studies
Lasso	Least Absolute Shrinkage and Selection Operator
LSTM	Long Short-Term Memory
MOS	Military Occupational Specialty
NBER	National Bureau of Economic Research
OPA	Office of People Analytics
OUSD(P&R)	The Office of the Under Secretary of Defense for Personnel and Readiness
PCA	principal component analysis
RNN	recurrent neural network
ROI	return on investment
RPM	Retention Prediction Model
SRB	Selected Retention Bonus
U.S.	United States
VA	Department of Veterans Affairs

This page is intentionally blank.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE (DD-MM-YY) XX-05-2020	**2. REPORT TYPE** Final	**3. DATES COVERED (From – To)**

4. TITLE AND SUBTITLE Leveraging Machine Learning in Defense Personnel Analyses	**5a. CONTRACT NO.** HQ0034-14-D-0001
	5b. GRANT NO.
	5c. PROGRAM ELEMENT NO(S).
6. AUTHOR(S) Julie Lockwood Alan Gelder Matthew Goldberg Jennifer Brooks George Prugh	**5d. PROJECT NO.**
	5e. TASK NO. BE-6-4311
	5f. WORK UNIT NO.
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Institute for Defense Analyses 4850 Mark Center Drive Alexandria, VA 22311-1882	**8. PERFORMING ORGANIZATION REPORT NO.** IDA Paper P-13174 Log: H 20-000152
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) OUSD (P&R), Military Personnel Policy Pentagon, Rm 5A734	**10. SPONSOR'S / MONITOR'S ACRONYM(S)** OUSD (P&R)
	11. SPONSOR'S / MONITOR'S REPORT NO(S).

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Leaders in government and industry are bombarded with opportunities to apply machine learning and other "big data" techniques. Reaping the benefits of these advances and avoiding missteps requires that leaders understand what these techniques can and cannot provide. Paired with illustrative examples from defense personnel research, we describe policy applications for several machine learning techniques, organized by the type of question asked: "what?", "how?" (or "why?"), and "what if?" This structure can be applied to nearly any topic area. "What" questions seek to describe—not explain—past or present conditions, or to forecast future conditions. Machine learning often produces highly accurate and detailed predictions. Answering "how" or "why" questions requires data from intentional or natural experiments to inform a model and permit hypothesis testing about causal relationships. Forward-looking "what if" hypothetical questions ask what will happen if certain conditions change. Analyzing "what if" questions requires a blend of causal analysis and structured forecasting. Machine learning techniques offer to dramatically expand researchers' ability to capture intricate dimensions of human decision making in complex models. The Institute for Defense Analyses offers recommendations to assist leaders in making effective and judicious use of these techniques.

15. SUBJECT TERMS

Machine Learning, Causality, Forecasting, Defense Personnel Policy, Retention

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NO. OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** David M. Percich
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	U	90	**19b. TELEPHONE NUMBER (Include Area Code)** (703) 693-2238

This page is intentionally blank.