CSTE Injury Data Science Competencies

CSTE Injury Surveillance Workgroup

Council of State and Territorial Epidemiologists • 2635 Century Parkway NE, Suite 700, Atlanta, GA 30345 • 770.458.3811 • www.cste.org

Table of Contents

Introduction

Funding Acknowledgement

This document was produced by the CSTE Injury Surveillance Workgroup with support from the Centers for Disease Control and Prevention (CDC) Cooperative Agreement Number NU38OT000297-03-00.

Background

In 2020, The CSTE Injury Surveillance Workgroup identified gaps in applied injury epidemiology workforce training and skills assessment. After exploring examples of existing competencies for epidemiologists and public health professionals, graduate school curricula, and recent federal, state, and local position descriptions, the workgroup began developing data science competencies targeted towards injury epidemiologists. A subgroup of six CSTE Injury Surveillance Workgroup members representing six state and local jurisdictions met regularly with CDC National Center for Injury Prevention and Control (NCIPC) subject matter experts and CSTE staff to develop a list of competencies broken down into six domains: Leadership, Project Management, Data Acumen, Data Analysis and Programming, Subject Matter Expertise, and Data Reporting and Dissemination. The subgroup also developed an accompanying Implementation Guide (page 9) that outlines recommendations for applying the Injury Data Science Competencies in jurisdictional health department teams.

Data science includes standard public health and epidemiology methods, but also new methods that are increasingly popular as big data and machine learning become more prevalent in public health datasets. NCIPC defines population-health data science as "a multidisciplinary approach combining traditional epidemiologic methods and contemporary computer science techniques, with a particular focus on large and complex data sources, to improve the measurement and prevention of injury and violence in communities." [1]

Applied epidemiology is a team-based field that has long benefited from collaborations among team members with different levels of experience and skillsets. The same is inevitably true as the pillars of data science become more integrated into public health practice. Injury epidemiology specifically uses a wide variety of large, complex data sources and multidisciplinary methods. As a result, there is a need for guidance on the specific skills that injury epidemiology teams should work toward in their team structure. [2]

As data science teams must master statistical analysis methods and various programming languages, it is important that results and findings are effectively communicated and developed into data driven recommendations, products, and resources. As described in the Harvard Business Review, "cross-disciplinary teams composed of members with varying talents who work in close proximity. Empathy, developed through exposure to others' work, facilitates collaboration among the types of talent. Work is no longer passed between groups; it's shared among them. A team approach—hardly new, but newly applied—can get data science operations over the last mile, delivering the value they've created for the organization." [3] As such, a cross-disciplinary team approach enables collaboration of members with complementary subject matter expertise to move through project management, data wrangling and analysis, design, and data reporting/surveillance in a more effective, efficient, and refined manner. Working in teams can preserve the scientific integrity and rigor of studies and reports and strengthens the epidemiology behind the data science. In state and local public health departments, it is critical to

note that teams can be fluid and consider the key employees within and across departments: epidemiologists, evaluators, IT staff, informaticians, program managers, etc. Note that contractors, full-, and part-time employees can make up these data science teams. Examples of key employees are included in the definitions below.

Definitions and Team Roles

Given the multidisciplinary nature of data science and the methods used by a cross-disciplinary data science team, there is some overlap among the following data science definitions and team roles. For example, an IT professional often fulfills the roles of database admin, computer scientist, and/or data manager, but they rely on the subject matter expertise of epidemiologists and statisticians.

Definitions

- **Data acumen**- the ability to understand data, make good judgments about and good decisions with data, and use data analysis tools responsibly and effectively.
- **Data analysis and programming**- the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making while utilizing multiple statistical and data analysis software.
- **Data governance**- The knowledge of ethics surrounding the data collection, use, analysis, and interpretation including data ownership, privacy & security, validity, anonymity, algorithmic biases.
- **Data stewardship** - the management and oversight of an organization's data assets to help provide business users with high-quality data that is easily accessible in a consistent manner.
- **Data visualization**- the practice of developing and executing systems for effective visual communication.
- **Data wrangling -** the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics; the seven steps are discovering, structuring, cleaning, enriching, validating, and publishing. [1]
- **Health equity**- a goal that every person in a society has the opportunity to "attain [their] full health potential" and no one is "disadvantaged from achieving this potential because of social position or other socially determined circumstances." When considered in context of collecting and presenting health data, this concept necessitates consideration of the data that is collected, how disparities are explained, the solutions recommended, and discussion of limitations of data and influences, such as institutional racism, that may not easily be represented in the data. In practice, this includes analyzing overall health outcomes (injuries) and behaviors among a population, assessing for differences in those outcomes and behaviors among population groups, and developing a communication (i.e., translation) plan to share these findings to inform decision makers' policy and practice. [4,5]
- **Informatics**- The systematic application of knowledge about systems that capture, manage, analyze, and use information to improve population health.
- **Machine learning** - the process of data analysis using iterative analytical model building that can often be automated. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, and make decisions without the assumptions of traditional statistical methods [6].

- **Population-based data**- data collected from an entire group of people within a specified geographic region or demographic sub-population.
- **Project management**- the practice of initiating, planning, executing, controlling, and completing the work of a team to achieve specific goals and meet specific success criteria at the specified time.
- **Natural language processing**- a field of Artificial Intelligence (AI) that gives the machines the ability to read, understand and derive meaning from human languages.
- **Subject matter expertise**- the level of knowledge of the organization, program, and strategies to inform project design and data analysis and keep the team focused on the end outcomes.

Possible Team Roles on the Data Science Team

- **Data analyst**- Executes the process of analyzing data, including assessing data using a range of statistical tools and programming techniques.
- **Data and analytics manager**- Manages a team of analysts and data scientists.
- **Data architect**- Creates templates for data management systems to integrate, centralize, protect, and maintain data source.
- **Data engineer**- Develops, constructs, tests, and maintains architectures such as databases and large-scale processing systems.
- **Data scientist**- Performs tasks related to data exploration, including quality control, quality assessment, validation, establishment of methods and study design, and identification of biases. Data scientists may be trained in epidemiology, statistics, and/or computer science or related field.
- **Database administrator**- Ensures the database is available to all relevant users and is well maintained and secured.
- **Epidemiologist**- Epidemiologists are formally trained to study "the distribution and determinants of health-related states or events in specified populations to control health problems." [7] Consider the functions performed by the individual rather than the job title.
- **Machine learning engineer**- Understands how to deploy and run systems in production and is trained in the challenges of supporting machine learning products.
- **Statistician**- Collects, analyzes, and interprets qualitative and quantitative data using statistical theories and methods. Statisticians are formally trained in statistical methods and probability theory.

Competencies and Performance Criteria

The following competencies are grouped under six key domains: Leadership, Project Management, Data Acumen, Data Analysis and Programming, Subject Matter Expertise, and Data Reporting and Dissemination. As described in the team roles section above, it is expected that members of multidisciplinary teams will likely possess skills and have competencies across multiple key themes. Because no one team member is likely to have mastery over all of these areas, the team model is necessary for achieving overarching injury data science competencies.

As in all public health work, implementation of these Data Science Team Competencies should be grounded in a health equity framework. [5]

Leadership

1. Leads teams to address a public health injury issue or agency issue
2. Establishes mission, strategy, and goals of the team
3. Define roles within the team
4. Implements program management principles and methodologies to achieve desired outcomes
5. Establishes relationships that promote trust and inclusion within and outside of the health department (e.g., health department office of informatics, Department of Corrections, Child Protective Services, nonprofit organizations, community service organizations, local businesses)
6. Identifies and addresses gaps in the training infrastructure in order to build injury workforce capacity
7. Builds consensus through increased understanding, compromise, and collaboration among team members and across department programs
8. Collaborates to build ongoing, trusted relationships with partners to implement processes and products
9. Selects effective frameworks for teams developing, testing, and implementing products and deliverables (e.g., Scrum framework, Plan-Do-Study-Act (PDSA cycle))

Project Management

1. Collaborates as a team to address a public health injury issue or agency issue
2. Applies project management principles and methodologies: scope, planning, assessment, quality and risk management, and team management ^
3. Demonstrates critical thinking to solve complex problems
4. Conducts evaluation and strategic planning activities, including needs assessments and development of logic models related to the intersection of the injury program and the agency
5. Applies understanding of agency/stakeholder purpose related to the project and ability to make decisions within that context, including sustainability

Data Management and Governance

1. Data Management
    a. Applies understanding of the data lifecycle^
    b. Develops data management plan^
    c. Applies understanding of data architecture, data types and data formats, data modeling and design, including related technologies^
    d. Applies understanding of how to set up and maintain data platforms
    e. Identifies and collects data in multiple formats^
    f. Prepares data structure, manipulate, transform, and clean data^
    g. Assesses data quality and validates datasets by applying techniques to deal with missing values, outliers, unbalanced data, and data normalization^
    h. Designs algorithms for accessing and analyzing large amounts of data^
    i. Applies understanding of cloud computing technologies and cloud powered services design for data infrastructure and data handling services in relation to data storage and version control^
    j. Uses relational and non-relational databases and data warehouse solutions to process structured and unstructured data^
    k. Works with a variety of data sources (e.g., electronic health records (EHRs))

l. Uses best practice approaches for linkage of multiple datasets
m. Executes principles and practices related to reproducibility and version control

2. Data Governance and Ethics
    a. Utilizes metadata, Proportional Integral Derivative (PID) controller, data registries, data factories^
    b. Complies with laws, data standards, compliance regulations, and cultural and ethical considerations relevant to data collection, management, communication, and responsible use of data and information^*
    c. Follows program ethics guidelines and principles when collecting and disseminating data to ensure respect for all persons, minimal risk of harm, maximized benefits, and fair distribution of risks and benefits among persons who provide data*
    d. Analyzes and communicates potential conflicts of interest*
    e. Applies knowledge of industry standards around the ethical boundaries when disseminating and using data*
    f. Analyzes the sensitive nature of cultural, political, and policy differences and their impact in the data, data analysis, and outputs*
    g. Incorporates applicable state and local privacy laws into data dissemination and uses to protect confidentiality/Personally Identifiable Information (PII)*
    h. Evaluates ethical practices to adapt communication style and techniques to culturally diverse situations*
    i. Applies data security and privacy principles, including destruction of records and HIPAA-compliant security measures.

Data Analysis and Programming

1. Applies knowledge of Foundational Statistics
    a. Demonstrates knowledge of sampling, probability theory, and probability distributions
    b. Performs descriptive and inferential statistics using a statistical programming language^
    c. Chooses appropriate research design
    d. Conducts and interprets mixed-methods analysis
    e. Conducts and interprets qualitative analysis
    f. Analyzes and interprets information, including meaningful patterns/processes
    g. Conducts population-based surveillance to determine rate, incidence, etc.
2. Applies non-traditional data surveillance sources and methods
3. Applies natural language processing/capturing information from text analysis (e.g., syndromic surveillance)
4. Conducts spatial analysis - geocoding, spatial statistics
    a. Communicates data via mapping (e.g. geocoding techniques)
    Performs cluster analysis to assess for spatial distributions of health outcomes
    b. Conducts ecological analysis
5. Data linkage
    a. Implements appropriate deterministic and probabilistic data linkage based on data sources
    b. Assesses software options and methods used to meet project needs
    c. Performs quality assurance (QA) on linkage projects

6. Programming
    a. Applies knowledge of statistical programming language (e.g. SAS, Python, and R)^
    b. Applies knowledge of database querying language (e.g. SQL)^
    c. Applies knowledge of database design (e.g. relational/structured vs. unstructured) and best practices for use
7. Machine Learning
    a. Differentiates among four types of machine learning approaches (e.g. supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning)
    b. Determines which machine learning approach is best suited for a specific application and problem
    c. Applies machine learning model testing and experimentation
    d. Utilizes machine learning technology, algorithms, and tools
    e. Evaluates machine learning models
    f. Implements machine learning natural language processing methods
    g. Applies neural network and deep learning algorithms
    h. Utilizes performance and accuracy metrics for model validation and selection
    i. Deploys and monitors a validated model in an operational environment

Subject Matter Expertise

1. Characterizes injury problems and health equity indicators
2. Forms hypotheses surrounding injury problems and health equity indicators
3. Describes limitations of data sources
4. Applies traditional injury surveillance methods (CDC Injury Indicators, CSTE Injury Surveillance Toolkit) and tools WISQARS, NVDRS, CDC WONDER, etc.) for data analysis
5. Explores novel or emerging injury surveillance data sources, methods and tools for data analysis (e.g., machine learning, novel data sources, etc.)
6. Utilizes appropriate data science methods to solve problems and make improvements in a public health/injury epidemiology context
7. Conducts literature and evidence-based reviews
8. Assesses surveillance data needs¶
9. Actively participates in peer learning opportunities

Data Reporting and Dissemination

1. Develops outputs and products to inform public health and programs
2. Disseminates data to inform and advise public health policies, programs and/or strategies
3. Provides critical explanation in data results to aid in interpretation and utility
4. Produces peer-reviewed publications
5. Identifies effective visualizations for an audience
6. Creates basic visualizations (e.g., graphs, tables, charts) using readily accessible tools (e.g., Microsoft Office)
7. Designs advanced visualizations using novel or emerging tools
8. Creates interactive data sharing dashboards and/or platforms
9. Communicates and translates findings effectively using appropriate vocabulary and level of detail tailored to the audience as well as keeping health equity (e.g., avoid stigmatizing a group or population with significant disparities)

10. Delivers results to stakeholders that are focused on actionable recommendations or priorities taking into account health equity
11. Gathers feedback to improve data quality and actionability

The CSTE Data Science Competencies were adapted from the following individual level competencies:

¶Adopted from the CDC/CSTE Applied Epidemiology Competencies [8]

^Adopted from CDC DSU Competencies [9]

*Adopted from the Draft CDC Future of Work, "Data Fluency" Recommendations [10]

Implementation Guide

The following section includes guidance on implementing the Injury Data Science Competencies in jurisdictional health department teams.

Identifying the Injury Data Science Team members

When building or identifying an Injury Data Science Team specific to public health department structures, several key items should be considered. A helpful first step may be to conduct a team member self-assessment and evaluation of competencies. This can help to identify existing skills and knowledge gaps. Once strengths and areas for improvement have been identified, a team structure and strategic succession plan will help guide the selection of specific data science team member roles, handling rotating members of the team, developing a team meeting schedule, and overarching goals for the team.

Self-assessment and evaluating competencies

CSTE recommends that health department injury data science teams each conduct a self-assessment to get a sense of each team member's individual strengths and needs. Team members should evaluate their own knowledge, comfort, and interest in each domain and meet to discuss with team leaders or as a group. This exercise can inform how to best designate data science tasks and build a strong foundation within the health equity framework.

The CSTE Injury Surveillance Workgroup adapted the CSTE Data Science Team Training (DSTT) Self-Assessment and recommends the following assessment to accompany the Injury Data Science Competencies (page 4). For each of the three questions under a given domain, the responses are assigned a value from 1 to 5. Then, the 3 questions are combined by summing to give a total score for the domain. Responses indicating lower self-rated knowledge, greater importance to the current job, and greater interest in learning more were assigned more points (e.g., the response to item 1 "I am not familiar with this domain at all" receives a score of 5 points, the response to item 2 "Not at all important" receives a score of 1 point). When comparing the totals for each domain, those domains with the highest score can be interpreted as priority areas for training topics. Individual or team learning goals may not be identified through this assessment as it strictly addresses the domain level, which likely too broad as a learning goal.

I. Leadership

Leadership refers to anyone in a position to influence and guide the data science team effectively—a formal title is not required. Leaders set direction, drive innovation, and engineer change. Competencies in this domain include but are not limited to:

- Leading teams to address a public health issue
- Establishing the mission, strategy, and goals of a team
- Building consensus among team members and across department programs

1. What is your current level of knowledge, skill, and/or ability in the domain of leadership?

- I am not familiar with this domain at all.
- I am a beginner in this domain.
- I am somewhat literate in this domain

- I am proficient in this domain
- I am an expert in this domain

2. How important is the domain of leadership to your current job?

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

3. How interested are you in developing your knowledge, skill, and/or ability in the domain of leadership?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

II. Project Management

This domain can be defined as: the practice of initiating, planning, executing, controlling, and completing the work of a team to achieve specific goals and meet specific success criteria at the specified time. Competencies in this domain include but are not limited to:

- Apply project management principles and methodologies: scope, planning, assessment, quality and risk management, and team management
- Demonstrate critical thinking to solve complex problems
- Collaborate as a team to address a public health injury issue or agency issue

1. What is your current level of knowledge, skill, and/or ability in the domain of project management?

- I am not familiar with this domain at all.
- I am a beginner in this domain.
- I am somewhat literate in this domain
- I am proficient in this domain
- I am an expert in this domain

2. How important is the domain of project management to your current job?

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

3. How interested are you in developing your knowledge, skill, and/or ability in the domain of project management?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

III. Data Management and Governance

This domain involves the ability to understand data, to make good judgments about and good decisions with data, and to use data analysis tools responsibly and effectively.

The competencies of **data management** include but are not limited to:

- Understanding of the data lifecycle
- Ability to develop data management plan
- Understanding of data architecture, data types and data formats, data modeling and
- design, including related technologies.
- Ability to manipulate, transform, and clean data

1. What is your current level of knowledge, skill, and/or ability in the domain of data management?

- I am not familiar with this domain at all.
- I am a beginner in this domain.
- I am somewhat literate in this domain
- I am proficient in this domain
- I am an expert in this domain

2. How important is the domain of data management to your current job?

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

3. How interested are you in developing your knowledge, skill, and/or ability in the domain of data management?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

Data governance & ethics can be defined as the knowledge of ethics surrounding the data collection, use, analysis, and interpretation including data ownership, privacy, security, validity, anonymity, and algorithmic biases. Competencies in this domain include but are not limited to:

- Understanding of the principles of data protection, backup, privacy, ethics and responsible use
- Evaluates and integrates ethics guidelines and principles when disseminating or using data

1. What is your current level of knowledge, skill, and/or ability in the domain of data governance & ethics?

- I am not familiar with this domain at all.
- I am a beginner in this domain.
- I am somewhat literate in this domain
- I am proficient in this domain
- I am an expert in this domain

2. How important is the domain of data governance & ethics to your current job?

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

3. How interested are you in developing your knowledge, skill, and/or ability in the domain of data governance & ethics?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

IV. Data Analysis and Programming

These terms might be what first comes to mind when you hear the phrase "data science". Data analysis and programming can be defined as the process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making while utilizing multiple statistical and data analysis software.

Data analysis and programming covers a wide range of skills. We will break this domain down into 6 subdomains.

Foundational Statistics

Some elements of this domain are:

- Demonstrate knowledge of sampling, probability theory, and probability distributions
- Ability to implement descriptive and inferential statistics using a statistical programming language
- Ability to use the scientific method and choose appropriate research design
- Mixed-methods analysis
- Qualitative analysis

1. What is your current level of knowledge, skill, and/or ability in the domain of foundational statistics?

- I am not familiar with this domain at all.
- I am a beginner in this domain.
- I am somewhat literate in this domain
- I am proficient in this domain
- I am an expert in this domain

2. How important is the domain of foundational statistics to your current job?

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

3. How interested are you in developing your knowledge, skill, and/or ability in the domain of foundational statistics?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

Syndromic Surveillance

Syndromic Surveillance uses electronic patient data to detect, understand, and monitor health events. By tracking symptoms of patients in emergency departments— before a diagnosis is confirmed—public health can detect unusual levels of illness to determine whether a response is warranted.

1. What is your current level of knowledge, skill, and/or ability in the domain of syndromic surveillance?

- I am not familiar with this domain at all.
- I am a beginner in this domain.
- I am somewhat literate in this domain
- I am proficient in this domain
- I am an expert in this domain

2. How important is the domain of syndromic surveillance to your current job?

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

3. How interested are you in developing your knowledge, skill, and/or ability in the domain of syndromic surveillance?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

Geospatial Analysis

Geographic or geospatial analysis can reveal patterns in data by incorporating the spatial or location element into an analysis. Competencies in this domain include geocoding data, using a Geographic Information System (such as ArcGIS), and building visualizations.

1. What is your current level of knowledge, skill, and/or ability in the domain of geospatial analysis?

- I am not familiar with this domain at all.
- I am a beginner in this domain.
- I am somewhat literate in this domain
- I am proficient in this domain
- I am an expert in this domain

2. How important is the domain of geospatial analysis to your current job?

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

3. How interested are you in developing your knowledge, skill, and/or ability in the domain of geospatial analysis?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

Data Linkage

Record linkage is the process of comparing records across datasets to identify individuals contained in both. Linkages can supplement or validate data across datasets and can identify duplicate records for the same individual within one dataset, a fundamental requirement for accuracy and validity of event counts in any disease registry.

1. What is your current level of knowledge, skill, and/or ability in the domain of data linkage?

- I am not familiar with this domain at all.
- I am a beginner in this domain.
- I am somewhat literate in this domain
- I am proficient in this domain
- I am an expert in this domain
-

2. How important is the domain of data linkage to your current job?

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

3. How interested are you in developing your knowledge, skill, and/or ability in the domain of data linkage?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

Programming

Elements of this domain include:

- Knowledge of statistical programming language like Python and R
- Knowledge of database querying language like SQL
- Knowledge of database design (relational/structured vs. unstructured) and best practices for use

1. What is your current level of knowledge, skill, and/or ability in the domain of programming?

- I am not familiar with this domain at all.
- I am a beginner in this domain.
- I am somewhat literate in this domain
- I am proficient in this domain
- I am an expert in this domain

2. How important is the domain of programming to your current job?

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

3. How interested are you in developing your knowledge, skill, and/or ability in the domain of programming?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

Machine Learning

Elements of this domain include:

- Differentiates among four types of machine learning approaches
- Determines which machine learning approach is best suited for a specific application and problem.
- Applies machine learning technology, algorithms, and tools
- Applies natural language processing methods
- Applies neural network and deep learning algorithms
- Utilizes performance and accuracy metrics for model validation and selection

1. What is your current level of knowledge, skill, and/or ability in the domain of machine learning?

- I am not familiar with this domain at all.
- I am a beginner in this domain.
- I am somewhat literate in this domain
- I am proficient in this domain
- I am an expert in this domain

2. How important is the domain of machine learning to your current job?

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

3. How interested are you in developing your knowledge, skill, and/or ability in the domain of machine learning?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

V. Subject Matter Expertise

This domain includes competencies that are specific to injury epidemiologists and public health practitioners. Competencies in this domain include but are not limited to:

- Characterizes injury problems and health equity indicators
- Applies traditional injury surveillance methods for data analysis
- Explores novel or emerging injury surveillance data sources, methods, and tools for data analysis

1. What is your current level of knowledge, skill, and/or ability in the domain of subject matter expertise?

- I am not familiar with this domain at all.
- I am a beginner in this domain.
- I am somewhat literate in this domain
- I am proficient in this domain
- I am an expert in this domain

2. How important is the domain of subject matter expertise to your current job?

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

3. How interested are you in developing your knowledge, skill, and/or ability in the domain of subject matter expertise?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

VI. Data Reporting and Dissemination

This domain involves communicating surveillance findings to stakeholders, partners, and the public. Competencies in this domain include but are not limited to:

- Communicate findings effectively
- Deliver actionable results to stakeholders
- Produce peer-reviewed publications
- Apply various software to visualize date
- Create interactive dashboards and/or platforms

1. What is your current level of knowledge, skill, and/or ability in the domain of data reporting and dissemination?

- I am not familiar with this domain at all.
- I am a beginner in this domain.
- I am somewhat literate in this domain
- I am proficient in this domain
- I am an expert in this domain

2. How important is the domain of data reporting and dissemination to your current job?

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

3. How interested are you in developing your knowledge, skill, and/or ability in the domain of data reporting and dissemination?

- None at all
- A little
- A moderate amount
- A lot
- A great deal

Recommendations for strategic planning, hiring new staff, and succession planning

Health department data science teams will benefit from developing strategic and succession plans. Strategic plans can be team or departmental focused. In some cases, it may be helpful for the overall health department to include a section in their overarching strategic plan focused on data science innovation and/or training. Inclusion of such elements in strategic plans can justify increased funding to data science programs. Even if the strategic plan is small-scale and relevant only to the data science team members, it can provide a valuable framework for goal setting and assigning benchmarks. The Resource List on page 28 provides links to strategic planning tools from the Association of State and Territorial Health Officials (ASTHO) and the National Association of City and County Health Officials (NACCHO).

Succession plans provide a systematic approach for organizing the future direction of a public health program. They improve operations as organizations change and enhance their stability as staff transition away from current or traditional roles and as new staff are hired. Succession plans should be linked to the program or department's strategic plan.[11] In the post-COVID-19 era, public health will see increases in funding, and it is critical for health departments to be prepared for changes to the workforce. The de Beaumont Foundation provides a change package on Retention and Succession Planning, and this document is linked in the Resource List below on page 28. Below are examples of position descriptions that can be adapted to reflect increasing needs for hiring epidemiologists, statisticians, and scientists with data science-related skills.

Example Position Descriptions

I. Computer Scientist (Data Scientist)

INTRODUCTION

The incumbent serves as a Computer Scientist performing data science work that requires application of quantitative and qualitative research and analytics, statistical analysis, and building high quality prediction systems integrated with public health surveillance systems and programs at the local, state and national levels that are structured or unstructured for: analysis; improved understanding and communication; development/visualization of new concepts, and/or processes that add value to health services delivery and the decision making process.

MAJOR DUTIES and RESPONSIBILITIES

Research, Analysis, and Data Evaluation (35%)

- Designs experiments, tests hypotheses, and builds scalable data science models. Conducts advanced data analysis and designs highly complex algorithms using artificial intelligence (e.g., machine learning) methods. Applies advanced statistical, data science, and predictive modeling techniques to build, maintain, and improve on multiple real-time decision systems.
- Leads discovery processes with stakeholders to identify business requirements and expected outcome. Makes strategic recommendations on data collection, integration, storage, access, analysis, and retention requirements incorporating business requirements and knowledge of best practices. This may include large structured and unstructured datasets.
- Identifies relevant data available, including internal and external data sources, leveraging new data collection processes such as smart meters and geo-location information, or social media and unstructured web-based data.
- Develops innovative and effective approaches to solve client's analytics problems and communicates results and methodologies. Validates analysis using scenario modeling. Identifies/creates the appropriate algorithm to discover patterns.
- Develops experimental design approaches to validate finding or test hypotheses. Assesses opportunities to enhance the qualification and assurance of the information to strengthen the use case. Defines the validity of information through automated systems and tools, how long the information is meaningful, and what other information is related.
- Utilizes patterns and variations in the volume, speed and other characteristics of data supporting the initiative, the type of data (e.g., images, text, clickstream or metering data) in predictive analysis. Leads the design and deployment of enhancements and fixes to systems as needed. Provides business metrics for overall projects to show contributions and improvements from monitoring of initial stages and through multiple iterations.

Technical Consultation (50%)

- Collaborates with CIO subject matter experts to select the relevant sources of information, which may include non-traditional datasets. Works with IT teams to support data collection, integration, storage, access, analysis, and retention requirements based on the surveillance data

collected. This may include large structured and unstructured datasets. Solves client analytics problems and communicates results and methodologies. Works in iterative processes with the client and validates findings.

- Works with stakeholders to identify the business requirements and the expected outcome. Works with and alongside CIO officials and subject matter experts by suggesting data science tools, methods, and statistical learning models of interest. Models and frames business scenarios that are meaningful and have impact on critical processes and/or decisions.
- Develops usage and access control policies and systems in collaboration with the IT security experts to ensure that the information used follows the compliance, access management, and control policies and that it meets CIO qualification and assurance requirements. Partners with researchers and subject matter experts in continuous improvement processes impacting data quality in the context of the specific use case. Recommends using innovative data science strategies for ongoing improvements to methods and algorithms that lead to findings, including new information.
- Maintains current knowledge of developments in statistical and machine learning analysis, and the use of technology to identify and evaluate new data science tools and methods to support current and future CIO efforts.

Data Presentation and Dissemination (15%)

- Presents and depicts rationale findings in layman's terms. Presents back results that contradict common assertions or beliefs, if needed. Communicates and works with business subject matter experts to educate the organization both from IT and business perspectives on new approaches, such as testing hypotheses and statistical validation of results. Helps the organization understand the principles and the math behind the process to drive organizational buy-in.
- Assists in preparing comprehensive reports that include discussion of substantive issues related to data science methodology for public health surveillance and related applications, assessment of the adequacy and quality of data used in the analyses, and explanation of the methods, results, and relevance to the health issue under study. Supports the preparation of scientific articles, technical reports, and data files for publication or other public dissemination. Produces static and interactive data visualizations to communicate findings to internal and external audiences. Prepares materials for presentations to other scientific staff and to federal, state, and local health program managers; and other health officials and health-related organizations.

Performs other duties as assigned.

KNOWLEDGE REQUIREMENTS

- Mastery of the principles, theories and practices of computer science and data science and the application of this field to the collection, processing, analysis, dissemination, storage and retrieval of health data to extend and adapt existing approaches, and to apply new developments to critical or obscure problems in public health data science.
- Mastery of advanced mathematical and statistical theories, principles, concepts, and practices, and skill in applying this knowledge to statistical procedures; complex mathematics; data or quantitative analysis; complex sample designs; data collection techniques; reduction of data

from multiple sources; computer techniques such as numerical and statistical analysis, probability theory, wave propagation modeling, and/or numerical simulation.

- Expert skill in developing problem-oriented and nonprocedural languages and their translating/operating systems, construction of input-output buffering schemes, and design of automatic scheduling and monitoring methods to increase scope and effectiveness of computer applications.
- Skill in researching computational complexities and the analysis of algorithms to explore data structures that lead to highly efficient combinatorial algorithms. Skill in developing abstract complexity theory to provide a theoretical foundation for understanding properties of the running times of efficient computer programs.
- Skill in advanced concepts of automation and information processing display, control, and transfer, programming, such as Hadoop MapReduce or other big data frameworks, Java), statistical modeling, and machine learning algorithms to advise on the analysis and interpretation of public health surveillance, research, and administrative data; to engage in research and data analysis; and to make presentations of findings to a variety of audiences.
- Broad knowledge of computer systems and hardware concepts such as computer architecture, computer communication systems, peripheral control systems, and bus architectures, cloud technology, statistical, machine learning applications and operations research, to apply available modeling and machine learning tools for collection, analysis, evaluation, reporting, and dissemination of health data that include structured and unstructured data, such as free text, and to evaluate the appropriateness and feasibility of using new and emerging technologies and applications in public health surveillance programs and other data development initiatives.
- Knowledge of program goals, the complexity of developing and conducting public health surveillance, appropriate uses of administrative data systems, and strategies for data analysis and dissemination, to include knowledge of public health surveillance design, data linkage, data validation and analysis processes, and confidentiality requirements, to function as an expert in public health statistics and machine learning research.
- Knowledge of and skill in applying research principles and practices, to plan, develop, conduct and report on research and analysis of significant issues and problems facing the public health community.
- Interpersonal skills, to establish and maintain effective working relationships with a variety of individuals and groups. Good written and oral communication skills, to prepare reports and manuscripts, and to make presentations of a variety of audiences.

II. Health Scientist (Data Scientist)

INTRODUCTION

The incumbent serves as a Health Scientist performing data science work that requires extraction of knowledge from public health surveillance systems and programs at the local, state and national levels that are structured or unstructured for: analysis; improved understanding and communication; development/visualization of new concepts, and/or processes that add value to health services delivery and the decision-making process.

MAJOR DUTIES and RESPONSIBILITIES

Research, Analysis, and Data Evaluation (35%)

- Consults and collaborates with statistical, data science, artificial intelligence (e.g., machine learning), and public health professionals in the collection, linkage, processing, coding, classification, and analysis of public health surveillance, research, and administrative health data.
- Learns and uses current data science methods and tools to plan and conduct research using public health data systems, including survey data, health care facility data, syndromic surveillance data, electronic health records, vital records, and non-traditional data sources such as social media and unstructured web-based data.
- Supports the development of proposals and projects that align with research and policy goals for data science research and analytic projects, defining the scope and intent of the projects, the data to be used, the analytic approaches and methods, technology resource requirements, timelines and significant milestones, and intended outputs.
- Assists in monitoring data quality issues as they relate to user data products and collaborates with Informatics and Information Technology professionals to develop automated systems and tools to process, clean, and verify data integrity. Assesses data elements and the implications of national data standards for the agency's data collection activities.
- Contributes to creating new data science analytic methods or modifies current methods (such as machine learning) and uses current technology to achieve the desired results while ensuring the scientific integrity and validity of the outputs. Participates in developing scalable data models and algorithms that can be applied to various topic areas.
- Provides analysis and supports data interpretation to both technical expert and lay users of public health data that include structured and unstructured data, such as free text, images, and data of mixed types. Contributes to ensuring that resulting products meet user needs and adhere to data quality standards.

Technical Consultation (50%)

- Brings and develops expertise in the fields of health science, artificial intelligence (e.g., machine learning), and programming languages as applied to public health. Serves as a data science expert in the analysis and classification of public health surveillance, research, and administrative health data.
- Collaborates with other professionals within and outside the Center in the conduct of surveillance, research, and analytical studies. Consults with Center epidemiologists, statisticians, computer scientists, economists, policy analysts, and public health professionals concerning ongoing and established studies or other projects where extensive analytic methodological support or innovation is required.
- Provides technical input to Center analysts and programmers in updating data from multiple sources with appropriate technical documentation. Collaborates with Informatics and Information Technology professionals to develop data storage and management systems that enable analysis of large structured and unstructured datasets.

- Provides advice on the use of data science tools, methods, and statistical learning models to collect, link, process, code, classify, and analyze public health surveillance, research, and administrative data. Assists in creating recommendation for additional research and development efforts and formulates proposals for new studies and data science projects related to the Center's priority topic areas. Supports synthesizing and interpreting the relevant literature and other public sources and provides analytical review of current methodological developments.
- Maintains current knowledge of developments in allied health sciences, modeling, and machine learning analysis, in conjunction with the use of technology to identify and evaluate new data science tools and methods to support current and future efforts.
- Participates as the expert in work groups focused on data science and machine learning as applied to public health data. Serves as for projects assigned. Participates in training and technical development activities.

Data Presentation and Dissemination (15%)

- Assists in preparing comprehensive reports that include discussion of substantive issues related to data science methodology for public health surveillance and related applications, assessment of the adequacy and quality of data used in the analyses, and explanation of the methods, results, and relevance to the health issue under study.
- Supports the preparation of scientific articles, technical reports, and data files for publication or other public dissemination. Produces static and interactive data visualizations to communicate findings to internal and external audiences. Prepares materials for presentations to other scientific staff and to federal, state, and local health program managers; and other health officials and health-related organizations.

Performs other duties as assigned.

Knowledge Required by the Position:

- Knowledge of the theories, principles, and concepts of epidemiology or related health sciences, to formulate hypotheses regarding health status and health outcomes, and to conduct meaningful analysis and interpretation of population health data from public health surveillance, research, and administrative data systems.
- Expertise in applying data science theory, concepts, principles, and methodology (including machine learning) to advise on the analysis and interpretation of public health surveillance, research, and administrative data; to engage in research and data analysis; and to make presentations of findings to a variety of audiences.
- Knowledge of program goals, the complexity of developing and conducting public health surveillance, appropriate uses of administrative data systems, and strategies for data analysis and dissemination, to include knowledge of public health surveillance design, data linkage, data validation and analysis processes, and confidentiality requirements, to function as an expert in public health statistics and machine learning research.

- Broad knowledge of computer systems, cloud technology, and statistical and machine learning applications, to apply available modeling and machine learning tools for collection, analysis, evaluation, reporting, and dissemination of health data that include structured and unstructured data, such as free text, and to evaluate the appropriateness and feasibility of using new and emerging technologies and applications in public health surveillance programs and other data development initiatives.
- Knowledge of and skill in applying research principles and practices, to plan, develop, conduct and report on research and analysis of significant issues and problems facing the public health community.
- Interpersonal skills, to establish and maintain effective working relationships with a variety of individuals and groups. Good written and oral communication skills, to prepare reports and manuscripts, and to make presentations of a variety of audiences.

III. Statistician (Data Scientist)

INTRODUCTION

The incumbent serves as a Statistician performing data science work that requires application of data mining techniques, statistical analysis, and building high quality prediction systems integrated with public health surveillance systems and programs at the local, state and national levels that are structured or unstructured for: analysis; improved understanding and communication; development/visualization of new concepts, and/or processes that add value to health services delivery and the decision making process.

MAJOR DUTIES and RESPONSIBILITIES

Research, Analysis, and Data Evaluation (35%)

- Designs experiments, tests hypotheses, and builds scalable data science models. Conducts advanced data analysis and designs complex algorithms using artificial intelligence (e.g., machine learning) methods. Identifies relevant data available, including internal and external data sources, leveraging new data collection processes such as smart meters and geo-location information, or social media and unstructured web-based data.
- Develops experimental design approaches to validate finding or test hypotheses. Validates analysis by comparing appropriate samples. Employs the appropriate data science algorithm to discover patterns. Assesses expected qualification and assurance of the information in support of the use case. Defines the validity of information through automated systems and tools, how long the information is meaningful, and what other information is related.
- Works with IT security experts to ensure that the information used is in compliance with the regulatory and security policies in place. Qualifies where information can be stored or what information, external to the organization, may be used in support of the use case.

- Identifies and analyzes patterns in the volume of data supporting the initiative, the type of data (e.g., images, text, clickstream or metering data) and the speed or sudden variations in data collection. Provides on-going tracking and monitoring of performance of decision systems and statistical models. Troubleshoots and implements enhancements and fixes to systems as needed. Provides business metrics for overall projects to show contributions and improvements from monitoring of initial stages and through multiple iterations.

Technical Consultation (50%)

- Collaborates with CIO subject matter experts to select the relevant sources of information, which may include non-traditional datasets. Works with IT teams to support data collection, integration, storage, access, analysis, and retention requirements based on the surveillance data collected. This may include large structured and unstructured datasets. Solves client analytics problems and communicates results and methodologies. Works in iterative processes with the client and validates findings.
- Works with stakeholders to identify the business requirements and the expected outcome. Works with and alongside CIO officials and subject matter experts by suggesting data science tools, methods, and statistical learning models of interest. Models and frames business scenarios that are meaningful and have impact on critical processes and/or decisions.
- Collaborates with the data steward to ensure that the information used follows the compliance, access management, and control policies and that it meets CIO qualification and assurance requirements. Partners with researchers and subject matter experts to define the data quality expectation in the context of the specific use case. Recommends using innovative data science strategies for ongoing improvements to methods and algorithms that lead to findings, including new information.
- Maintains current knowledge of developments in statistical and machine learning analysis, and the use of technology to identify and evaluate new data science tools and methods to support current and future CIO efforts.

Data Presentation and Dissemination (15%)

- Presents and depicts rationale findings in layman's terms. Presents back results that contradict common assertions or beliefs, if needed. Communicates and works with business subject matter experts to educate the organization both from IT and business perspectives on new approaches, such as testing hypotheses and statistical validation of results. Helps the organization understand the principles and the math behind the process to drive organizational buy-in.
- Assists in preparing comprehensive reports that include discussion of substantive issues related to data science methodology for public health surveillance and related applications, assessment of the adequacy and quality of data used in the analyses, and explanation of the methods, results, and relevance to the health issue under study.

- Supports the preparation of scientific articles, technical reports, and data files for publication or other public dissemination. Produces static and interactive data visualizations to communicate findings to internal and external audiences. Prepares materials for presentations to other scientific staff and to federal, state, and local health program managers; and other health officials and health-related organizations.

Performs other duties as assigned.

Knowledge Required by the Position:

- Knowledge of the theories, principles, and concepts of statistical techniques, such as statistical
- packages, statistical analysis, quantitative analytics, forecasting/predictive analytics, multivariate testing, optimization algorithms, and data science methods such as machine learning to provide solutions to loosely defined problems through leveraging pattern detection over large datasets.
- Skill in programming, such as Hadoop MapReduce or other big data frameworks, Java), statistical modeling, and machine learning algorithms to advise on the analysis and interpretation of public health surveillance, research, and administrative data; to engage in research and data analysis; and to make presentations of findings to a variety of audiences.
- Knowledge of program goals, the complexity of developing and conducting public health surveillance, appropriate uses of administrative data systems, and strategies for data analysis and dissemination, to include knowledge of public health surveillance design, data linkage, data validation and analysis processes, and confidentiality requirements, to function as an expert in public health statistics and machine learning research.
- Broad knowledge of computer systems, cloud technology, and statistical and machine learning applications, to apply available modeling and machine learning tools for collection, analysis, evaluation, reporting, and dissemination of health data that include structured and unstructured data, such as free text, and to evaluate the appropriateness and feasibility of using new and emerging technologies and applications in public health surveillance programs and other data development initiatives.
- Knowledge of and skill in applying research principles and practices, to plan, develop, conduct and report on research and analysis of significant issues and problems facing the public health community.
- Interpersonal skills, to establish and maintain effective working relationships with a variety of individuals and groups. Good written and oral communication skills, to prepare reports and manuscripts, and to make presentations of a variety of audiences.

Connecting knowledge and access/application of the competencies

In order for the data science competencies to be of utility to a data science team, the knowledge and skillsets of these competencies must be translated into action through the application of skills. The self-evaluation of competencies as well as the structure and strategic plan for a data science team will help team members to understand how to best use their skills and apply them as a member of the broader team. Team members should discuss how the competencies can be useful to both old and new applications within their jurisdictions, including knowledge sharing with other departments and jurisdictions to further develop the competency focus areas.

Resource List

Data Science for Injury and Violence Prevention	https://www.cdc.gov/injury/pdfs/data-science/Data-Science-Strategy_FINAL_508.pdf
The Data Science Skills Competency Model: A blueprint for the growing data scientist profession	https://www.ibm.com/downloads/cas/7109RLQM
Data Science and the Art of Persuasion	https://hbr.org/2019/01/data-science-and-the-art-of-persuasion
Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings	https://academic.oup.com/jamia/article/26/5/383/5369358
Data acumen definition	https://www.nap.edu/read/25104/chapter/4
What is project management	https://www.pmi.org/about/learn-about-pmi/what-is-project-management
The different data science roles in the industry	https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html
15 habits I stole from highly effective data scientists	https://towardsdatascience.com/15-habits-i-stole-from-highly-effective-data-scientists-441b1d46c572
Your guide to natural language processing	https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1
CDC Health Equity Style Guide	https://ehe.jhu.edu/DEI/Health_Equity_Style_Guide_CDC_Reducing_Stigma.pdf
Public Health Foundation Competency Rules	http://www.phf.org/programs/corecompetencies/Pages/Core_Competencies_Domains.aspx
ASTHO Strategic Planning Guide: Guidance and Resources to Assist State and Territorial Health Agencies in Developing a Strategic Plan	https://www.astho.org/Accreditation-and-Performance/Strategic-Planning-Guide/Home/
NACCHO Strategic Planning	https://www.naccho.org/programs/public-health-infrastructure/performance-improvement/strategic-planning
de Beaumont Foundation Change Package: Retention and Succession Planning	https://debeaumont.org/wp-content/uploads/2019/04/R5_SuccessionPlanning.pdf

References

1. Data Science Strategy for Injury and Violence Prevention. Published online August 2020. https://www.cdc.gov/injury/pdfs/data-science/Data-Science-Strategy_FINAL_508.pdf

2. Mew M. Data Scientists Will be Extinct in Ten Years: And Why It's Not a Bad Thing. Published online May 10, 2021. https://towardsdatascience.com/data-scientists-will-be-extinct-in-10-years-a6e5dd77162b

3. Berinato S. Data Science and the Art of Persuasion. Published online February 2019. https://hbr.org/2019/01/data-science-and-the-art-of-persuasion

4. Health Equity. Published online March 11, 2020. https://www.cdc.gov/chronicdisease/healthequity/index.htm

5. Health Equity Style Guide for the COVID-19 Response: Principles and Preferred Terms for NonStigmatizing, Bias-Free Language. Published online August 11, 2020.

6. The Data Science Skills Competency Model: A blueprint for the growing data scientist profession. Published online January 2020. https://www.ibm.com/downloads/cas/7109RLQM

7. Last JM, International Epidemiological Association, eds. *A Dictionary of Epidemiology*. 4th ed. Oxford University Press; 2001.

8. Applied Epidemiology Competencies. Published online February 2008. http://www.cste2.org/webpdfs/AppliedEpiCompwcover.pdf

9. Personal communication on August 13, 2021, CDC DSU Competencies.

10. Personal communication on August 13, 2021,CDC Future of Work report, "Data Fluency" Recommendations.

11. Wiesman, John MPH, CPH; Baker, Edward L. MD, MPH Succession Planning and Management in Public Health Practice, Journal of Public Health Management and Practice: January/February 2013 - Volume 19 - Issue 1 - p 100-101 doi: 10.1097/PHH.0b013e318272bb09

12. Dichev, C. & Dicheva, D. (2017). Towards data science *Procedia Computer Science, 108,* 2151-2160.

13. Obi Tayo B. Data Science Minimum: 10 Essential Skills You Need to Know to Start Doing Data. Published online October 2020.