







## SHEA Position Paper

# SHEA position statement on pandemic preparedness for policymakers: pandemic data collection, maintenance, and release

Westyn Branch-Elliman MD, MMSc<sup>1,2,3</sup> , David B. Banach MD, MPH, MS<sup>4,5</sup> , Lynne J. Batshon BS<sup>6</sup>, Ghinwa Dumyati MD<sup>7,8</sup> , Sarah Haessler MD<sup>9,10</sup>, Vincent P. Hsu MD, MPH<sup>11,12</sup>, Robin L.P. Jump MD, PhD<sup>13,14</sup>, Anurag N. Malani MD<sup>15</sup>, Trini A. Mathew MD, MPH<sup>16,17,18,19</sup>, Rekha K. Murthy MD<sup>20,21</sup>, Steven A. Pergam MD, MPH<sup>22,23,24</sup> , Erica S. Shenoy MD, PhD<sup>3,25,26</sup>  and David J. Weber MD, MPH<sup>27</sup> 

<sup>1</sup>Veterans Affairs Boston Healthcare System, Boston, MA, USA, <sup>2</sup>VA National Artificial Intelligence Institute (NAII), Washington, DC, USA, <sup>3</sup>Harvard Medical School, Boston, MA, USA, <sup>4</sup>University of Connecticut School of Medicine, Farmington, CT, USA, <sup>5</sup>Yale School of Public Health, New Haven, CT, USA, <sup>6</sup>Society for Healthcare Epidemiology of America (SHEA), Arlington, VA, USA, <sup>7</sup>University of Rochester Medical Center, Rochester, NY, USA, <sup>8</sup>Center for Community Health, Rochester, NY, USA, <sup>9</sup>Baystate Medical Center, Springfield, MA, USA, <sup>10</sup>University of Massachusetts Chan Medical School – Baystate, Springfield, MA, USA, <sup>11</sup>AdventHealth, Altamonte Springs, FL, USA, <sup>12</sup>Loma Linda University School of Medicine, Loma Linda, CA, USA, <sup>13</sup>Geriatric Research Education and Clinical Center (GRECC), Veterans Affairs Pittsburgh Healthcare System, Pittsburgh, PA, USA, <sup>14</sup>University of Pittsburgh School of Medicine, Pittsburgh, PA, USA, <sup>15</sup>Trinity Health Michigan, Ann Arbor, MI, USA, <sup>16</sup>HealthTAMCycle3, PLLC, Troy, MI, USA, <sup>17</sup>Corewell Health, Taylor, MI, USA, <sup>18</sup>School of Medicine, Wayne State University, Detroit, MI, USA, <sup>19</sup>Oakland University William Beaumont, Rochester, MI, USA, <sup>20</sup>Cedars-Sinai, Los Angeles, CA, USA, <sup>21</sup>David Geffen School of Medicine at UCLA, Los Angeles, CA, USA, <sup>22</sup>Fred Hutchinson Cancer Research Center, Seattle, WA, USA, <sup>23</sup>University of Washington, Seattle, WA, USA, <sup>24</sup>Seattle Cancer Care Alliance, Seattle, WA, USA, <sup>25</sup>Massachusetts General Hospital, Boston, MA, USA, <sup>26</sup>Mass General Brigham, Boston, MA, USA and <sup>27</sup>University of North Carolina, Chapel Hill, NC, USA

## Abstract

The Society for Healthcare Epidemiology in America (SHEA) strongly supports modernization of data collection processes and the creation of publicly available data repositories that include a wide variety of data elements and mechanisms for securely storing both cleaned and uncleaned data sets that can be curated as clinical and research needs arise. These elements can be used for clinical research and quality monitoring and to evaluate the impacts of different policies on different outcomes. Achieving these goals will require dedicated, sustained and long-term funding to support data science teams and the creation of central data repositories that include data sets that can be “linked” via a variety of different mechanisms and also data sets that include institutional and state and local policies and procedures. A team-based approach to data science is strongly encouraged and supported to achieve the goal of a sustainable, adaptable national shared data resource.

(Received 9 February 2024; accepted 16 February 2024; electronically published 5 June 2024)

## Background

Without a centralized, national data system, the public health community persistently encountered gaps in data availability during the COVID-19 pandemic, delaying analysis and scientific advancements essential for informing and updating pandemic policy response recommendations. This commentary provides recommendations for the creation and maintenance of such a centralized data system to improve the nation’s ability to detect and iteratively respond to matters important to public health, including future pandemics, in a data-informed manner.

## Rationale

Moving forward, the creation of a national data repository with a wide variety of data elements and types is essential to develop and

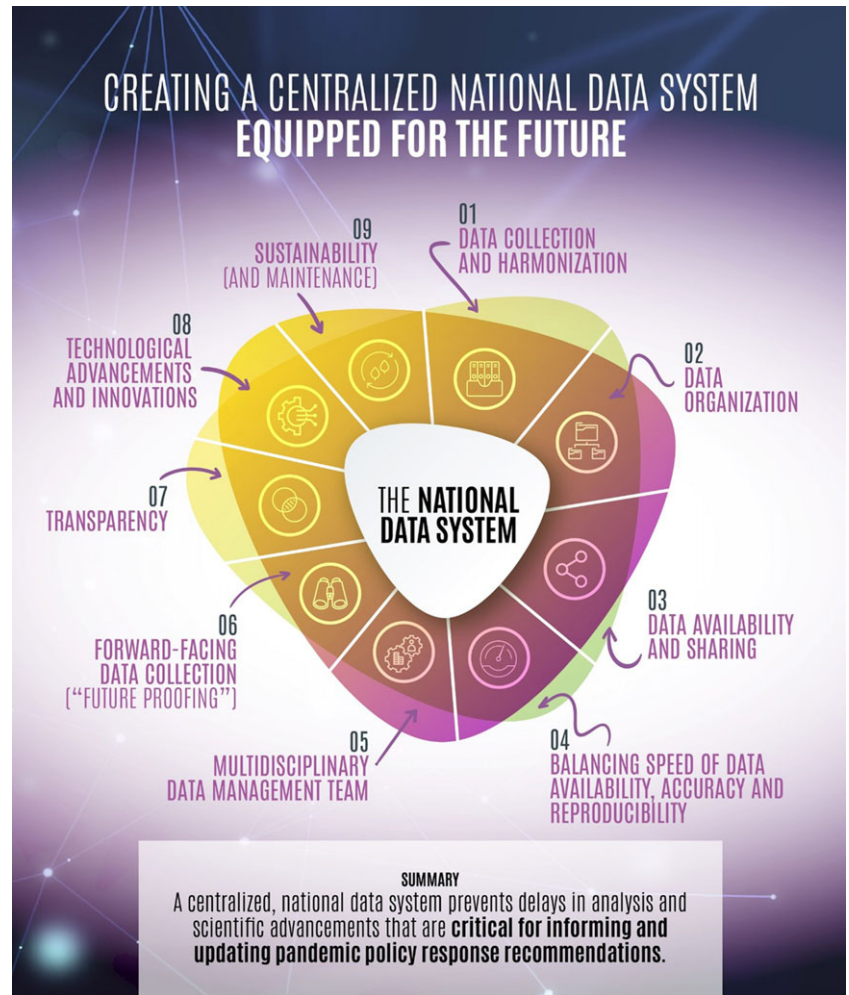
inform future practice and policy. Efforts to modernize data management and sharing should balance speed, accuracy, reliability, transparency, and feasibility (Figure 1). They should include strategies to “future proof” data, as forecasting the necessary data elements is not always possible. Agile data systems can help answer research and policy questions. Thus, data collection systems must be forward-thinking and continuously adaptable, expandable, and linkable to meet current and future needs.

## Recommendations

Modernizing data management and access is a top priority for the *Society for Healthcare Epidemiology in America* (SHEA). A centralized, national data system should focus on secure data collection, harmonization, and curation with dedicated teams of data scientists and clinical experts. Ideally, the system should be searchable with rapid access to data that can be de-identified. The ability to retain identifiers in a secure national repository is necessary for future linkage to other data sets or other data elements. Historically, most national, centralized data systems,

**Corresponding author:** Westyn Branch-Elliman; Email: [wbranch@bidmc.harvard.edu](mailto:wbranch@bidmc.harvard.edu)

**Cite this article:** Branch-Elliman W, Banach DB, Batshon LJ, et al. SHEA position statement on pandemic preparedness for policymakers: pandemic data collection, maintenance, and release. *Infect Control Hosp Epidemiol* 2024. 45: 821–825, doi: [10.1017/ice.2024.65](https://doi.org/10.1017/ice.2024.65)



**Figure 1.** Key elements of a modernized public health data system to support future infectious diseases public health challenges.

such as the Centers for Disease Control and Prevention's (CDC) National Healthcare Safety Network, have focused primarily on the collection of quantitative variables. However, an advanced data sharing and management system should encompass a wide variety of data types in anticipation for technological advancements capable of extracting data from sources currently inaccessible. The data repository should include both qualitative and quantitative data elements to promote innovation and policy evaluation. Considerations of different features and principles of a modernized data management and sharing system follow. Specific recommendations to policymakers are presented in Table 1.

#### *Data collection and harmonization*

Modernizing data collection with the creation of a shared national data resource will allow for near-real time data analysis and provide a platform for future technological advancements and improvement in the practice of early detection and infection prevention.

Data platforms must have advanced mechanisms for securing and storing data to protect individually identifiable health information, similar to protections employed in other sectors, such as the banking industry.

#### *Data organization*

Recognizing that many scientific questions are not clear in advance, national data sets with detailed cross-referencing ontology should

be created. These data sets must clearly explain how different data elements relate to each other and are organized, such that different data sources can be accurately and efficiently connected.

Data will come from a variety of sources, including state and local health departments and individual healthcare facilities, and identifiers need to be available for cross-linkage. Curated data elements should be organized. A data dictionary, with clear and reproducible definitions, should be available for review.

Challenges with linking different data sets and data elements has limited advancements during the pandemic. The removal of identifiers from data elements is permanent, meaning de-identified data elements in one data set can no longer be linked data elements in another data set to allow for accurate, granular analysis. In the absence of rich data sets, data collected at the individual, facility/healthcare system, community, state, or national level is often augmented with other data sets. Aggregated data is often insufficiently precise to allow for the granular, detailed analysis needed to inform clinical care and infection prevention decisions. Careful data management planning is required to ensure that relevant identifiers that can be used to cross-reference and link different data sets and data sources are maintained with important cybersecurity measures, as exist in other sectors. Failure to accurately and precisely match various data elements can lead to challenges with data analysis and interpretation, or, under the worst circumstances, uninterpretable findings. To facilitate innovation and accurate and actionable analysis, national data

**Table 1.** Pandemic Data Management: Challenges, Recommendations to Policymakers, and Examples

Challenge	Recommendations ( <i>examples</i> )
Variable data collection and harmonization	<ul style="list-style-type: none"> <li>▪ Expand data collection to focus on gathering the widest variety of data elements from the widest variety of sources:               <ul style="list-style-type: none"> <li>○ <i>Include qualitative and quantitative data elements.</i></li> <li>○ <i>Collect and organize data not traditionally included in organized data repositories, such as image data.</i></li> </ul> </li> <li>▪ Engage other sectors to develop data transfer and storage systems that are secure, with advanced mechanisms for storing and sharing data.</li> <li>▪ Explore secure cloud-based systems.</li> </ul>
Variable data organization	<ul style="list-style-type: none"> <li>▪ Develop and implement national standards for data organization, such as:               <ul style="list-style-type: none"> <li>○ <i>Cross-referencing ontology, with clear maps detailing how different data elements relate to each other and how they are organized.</i></li> <li>○ <i>Collecting at the smallest unit possible so that analyses not prespecified in advance can be completed.</i></li> <li>○ <i>Creating identifiers similar to county Federal Information Processing Standard (FIPS) codes that can be used for linkage and future analysis.</i></li> <li>○ <i>Using a variety of different data elements, including qualitative and quantitative data elements.</i></li> </ul> </li> </ul>
Limited data availability and sharing	<ul style="list-style-type: none"> <li>▪ Provide public access for data set download and analysis, while ensuring privacy protections for personal and healthcare system or facility-level identifiers.</li> <li>▪ Ensure accessible and transparent data to those potentially impacted by it, with a mechanism for review and update to promote accuracy, accountability, and equity.</li> <li>▪ Allow sharing of locally developed policies and procedures that may be helpful for other healthcare systems to include in the repository and maximize access.</li> </ul>
Difficulty in balancing speed of data availability, accuracy, and reproducibility	<ul style="list-style-type: none"> <li>▪ Integrate resiliency into the system, such that data scientists can pivot to pandemic responses and pandemic data management as needed.</li> <li>▪ Review necessary cleaning steps required for data release to reduce delays in data access and availability.</li> <li>▪ Ensure data elements such as facility and healthcare system procedures and protocols are available in a repository without a focus on cleaning or organization before release and accessibility.</li> </ul>
Outdated data systems	<ul style="list-style-type: none"> <li>▪ Pursue active data collection from healthcare systems and local/state public health to create a national data repository.</li> <li>▪ Ensure database adaptability to many common programming languages and in many common file formats.</li> <li>▪ Design data collection processes to gather the most amount of data possible, as we cannot always forecast what data will be needed to answer what questions in the future.</li> </ul>
Limited data transparency	<ul style="list-style-type: none"> <li>▪ Allow data access as widespread as possible to the extent feasible with HIPAA and privacy protection.</li> <li>▪ Code underlying database creation, linkage, and/or analysis for immediate public review and use.</li> </ul>
Lack of sustainable funding	<ul style="list-style-type: none"> <li>▪ Fund and support the creation of multidisciplinary data management teams.</li> <li>▪ Fund and support improved interfaces with healthcare systems to ease data sharing and collection.</li> <li>▪ Fund and support ongoing updates and re-evaluations of data systems.</li> <li>▪ Integrate cross-training and resiliency into the system.</li> </ul>

sets should be created that can interface with different electronic medical records and include multiple identifiers and the most granular unit of analysis possible that also maintains anonymity and Health Insurance and Portability and Accountability Act (HIPAA)-protections.

### **Data availability and sharing**

In line with the Open Science Framework,<sup>1</sup> the modernized data resource should include mechanisms for broad access to shared national resources to democratize data availability.

To the extent feasible, data sets should be made publicly available. If person or facility/healthcare system-level identifiers are required, a simple data use agreement process, with protections in place, could be used to protect human and facility subjects for information in which identifiers are necessary or complete deidentification is not possible, similar to the process used for the CDC restricted access COVID-19 database.<sup>2</sup> This process balances access and data safety and provides a mechanism for transferring and sharing large data sets. Once a sufficient number of cases have accrued to protect patient and/or facility/healthcare system privacy, anonymized data should be made widely available with minimal access restrictions.

### **Balancing speed of data availability and accuracy and reproducibility**

Mechanisms for improving the speed and accuracy of release of centrally collected data must be supported and advanced.

Creation of a national database with consistent and comparable data elements is challenging, as every healthcare system and every state and local government agency collects and transmits data in a different way. Thus, under current processes, data collection and availability necessitate substantial data cleaning before it can be analyzed or published for review. The time-consuming nature of data cleaning requires a balance between substance and efficiency. In some cases, data cleaning to ensure accuracy and reproducibility can lead to years long delay in data availability, as is the case for the US Renal Data System, which tracks nation-wide kidney disease and dialysis use.<sup>3</sup> To address the inherent tension between the speed of data availability (and apparent government transparency), a centralized data collection and management service could prioritize organization and release of a prespecified set of objective, quantitative data elements with a focus on early data release. Other data elements that are more difficult to clean and classify could be collected but not prioritized for cleaning, organization, and release. If future needs for these data elements arise, then the information could be organized by

data scientists with relevant expertise depending on the data type and made available. Such a system might balance future-proofing the database with speed of data availability.

### **Multidisciplinary data management team**

To achieve these data collection and management goals, we encourage the development of multidisciplinary teams, with expertise in computer science, informatics, mixed methods, qualitative data analysis, and clinical expertise including in infectious diseases and infection prevention with dedicated support to manage and continually adapt and improve national data sources.

Multidisciplinary data management teams, with dedicated funding and support, should focus on data procurement, storage, management, and organization to create a national data resource that can be broadly used.

### **Forward-facing, not backward-looking, data collection**

A future-proof national data repository with a variety of different data types and elements should be created and centrally managed and funded.

Accurately forecasting data elements that will be needed and useful in the future is not possible and complicates the creation of a national data repository. Observational research investigations are often plagued by missing data. To address this perennial research challenge, which limits the quality and interpretability of observational data sets, a variety of data sources should be collected and centrally stored. However, not all data elements need to be cleaned and then made publicly accessible. For example, facility-wide methicillin-resistant *Staphylococcus aureus* (MRSA) control policies could be collected and stored in a centralized data repository but not organized or categorically coded. If questions arise about the effectiveness of different bundles, unstructured data collected in the centralized data repository could be organized and analyzed at a future time point to resolve clinically important questions about policy real-world effectiveness. These structured and unstructured data elements should be available via a relatively streamlined data use agreement process to balance access with protection of participating parties. If access is granted to these centrally collected materials, coders should then provide their coding scheme, definitions, and output to the centralized data repository for future use and validation.

To achieve these goals, the national data repository should actively seek facility/healthcare system and state and local data and facilitate data collection, rather than continuing to pursue a passive, voluntary data collection process. Recognizing that many institutions and state and local governments have limited resources to support data sharing, national resources including federal funding and personnel support should be dedicated to assisting with these efforts. For state-level data, participation should be strongly encouraged and supported. Common and consistent definitions and data elements, including elements that can be used to evaluate health equity outcomes, should be applied and collected to allow comparisons. To encourage participation, data included in the data repository should be readily accessible to those willing to share with others.

### **Transparency**

Lowering of public trust fueled by limited transparency is an ongoing public health crisis that must be addressed and alleviated through increasing data access and transparency of analysis.

Lack of trust has complicated public health efforts during the pandemic.<sup>4,5</sup> Coupled with advancements in health communications,<sup>6</sup> more transparent data access may help to improve trust in governmental institutions and in scientific expertise and improve uptake of public health policy recommendations.<sup>7</sup> To promote trust, reliability, and reproducibility, code underlying database creation, linkage, and any analysis should be available for review and comment with feedback.

### **Technological advancements and innovations**

To improve early warning systems and to facilitate future pandemic responses, data repositories must be modernized and standardized.

Modernized data repositories are an essential first step toward leveraging emerging technologies, such as machine learning and other artificial intelligence-based tools, to improve early warning systems and to facilitate pandemic responses. To support future advancement, databases should be adaptable to many common programming languages and in many common file formats.

### **Sustainability**

To achieve innovation in infection detection and pandemic preparedness, centralized data management systems need to be sustainable for the long term, with a dedicated and nationally funded and supported data management team.

The data management team should have dedicated funding and multidisciplinary expertise and should focus on creation and maintenance of a national repository as its primary goal and purpose. Specific assignments of the study team should include data collection, quality control and review, data harmonization and curation, and steps to de-identify and release data. The data science team should also ensure that data are easily organized and available to programmers and analysts for use and review. Sustainability can also be attained by integrating and supporting an interface with healthcare system, state, and local data collection systems, to reduce the need for data cleaning, and by providing external facilitation to smaller organizations to help them make data available. The dedicated data science team should be flexible and able to pivot as needed to mitigate an evolving public health emergency.

### **Summary**

In summary, SHEA strongly supports modernization of data collection processes and the creation of publicly available data repositories that include a wide variety of data elements and mechanisms for securely storing both cleaned and uncleaned data sets that can be curated as clinical and research needs arise. These elements can be used for clinical research and quality monitoring and to evaluate the impacts of different policies on different outcomes. Achieving these goals will require dedicated, sustained, and long-term funding to support data science teams and the creation of central data repositories that include data sets that can be “linked” via a variety of different mechanisms and also data sets that include institutional and state and local policies and



procedures. A team-based approach to data science is strongly encouraged and supported to achieve the goal of a sustainable, adaptable national shared data resource.

**Acknowledgments.** The authors would like to thank the VA National Artificial Intelligence Institute for assistance with graphic design.

**Financial support.** None.

**Competing interests.** WBE reports research funding from the Health Services Research and Development Service (US Department of Veterans Affairs) and research support from Gilead Sciences (Funds to Institution). She also reports salary support from the VA National Artificial Intelligence Institute.

**Disclaimer.** The views presented are those of the authors on behalf of the Society for Healthcare Epidemiology in America and do not necessarily represent those of the US Department of Veterans Affairs or the US Federal Government.

## References

1. Foster ED, Deardorff A. Open science framework (OSF). *Journal of the Medical Library Association: JMLA* 2017;105:203.
2. Centers for Disease Control and Prevention. *COVID-19 Case Surveillance Restricted Access Detailed Data*. Atlanta, GA: Centers for Disease Control and Prevention; 2023.
3. National Institute of Diabetes and Digestive and Kidney Disease. *United States Renal Data System Progress through Research*. Bethesda, MD: National Institute of Diabetes and Digestive and Kidney Disease; 2023.
4. Bollyky TJ, Hullah EN, Barber RM, *et al.* Pandemic preparedness and COVID-19: an exploratory analysis of infection and fatality rates, and contextual factors associated with preparedness in 177 countries, from Jan 1, 2020, to Sept 30, 2021. *The Lancet* 2022;399:1489–1512.
5. Perry J. *Trust in Public Institutions: Trends and Implications for Economic Security*. New York: United Nations Department of Economic and Social Affairs; 2021.
6. Steel-Fisher GK, Findling MG, Caporello HL, *et al.* Trust In US Federal, State, and Local Public Health Agencies during COVID-19: responses and policy implications: study reports the results of a survey of public trust in us federal, state, and local public health agencies' performance during the COVID-19 pandemic. *Health Affairs* 2023;42:328–337.
7. Kennedy B, Tyson A, Funk C. *Americans' Trust in Scientists, Other Groups Declines*. Washington, DC: Pew Research Center; 2022.