*Proceedings of the 8th World Congress on New Technologies (NewTech'22)*
*Prague, Czech Republic– August 03-05, 2022*
*Paper No. ICBB 025*
*DOI: 10.11159/icbb22.025*

# An Analysis of Discrepancies in Commonly Used Measures of Autism Prevalence

**Aidan L. Lin[1], Sujata K. Bhatia[2]**
[1]Brophy College Preparatory, 4701 N Central Ave, Phoenix AZ, 85012. USA.
alin23@brophybroncos.org
[2]Harvard University, 51 Brattle St, Cambridge MA, 02138. USA.
sbhatia@g.harvard.edu

***Abstract*** **-** Autism Spectrum Disorder (ASD) is a developmental disability of variable severity that is characterized by challenges with social skills, repetitive behaviours, speech, and nonverbal communication. Alarming increases in the prevalence of autism spectrum disorder in the United States have been reported. Currently, there are four commonly used data sources (i.e., Special Education Child Count, National Survey of Children's Health, Medicaid, Autism and Developmental Disabilities Monitoring Network) related to ASD prevalence that are presented on the Centers for Disease Control and Prevention (CDC) database. However, the data availability and coverage, relative consistency, and accuracy of those data sources in the past 20 years are yet to be investigated. In this study, I quantitatively assessed the discrepancies of autism prevalence among the four CDC measures for the entire United States and its 50 states. Statistical analysis showed that there was a significant disparity among different measures. The data coverage, methodology, and design criteria for each measure were then investigated, with the recommendation of improving areas for each measure provided. It was evident that there was a high unmet need to reduce disparities in the identification of ASD, which could guide the collection of appropriate and reliable ASD prevalence data and support relevant scientific studies on possible causes and effective interventions.

***Keywords*****:** Autism Spectrum Disorder (ASD), autism prevalence, autism diagnosis, CDC data source

## 1. Introduction

Autism spectrum disorder (ASD) is a set of neurodevelopmental disorders characterized by a lack of social interaction, and verbal and nonverbal communications. The distinctive social behaviours include an avoidance of eye contact, problems with emotional control or understanding the emotions of others, and a markedly restricted range of activities and interests [1]. There is not just one but many autism subtypes, and each person with autism can have unique strengths and challenges. A combination of genetic and environmental factors influences the development of autism, and autism is often accompanied by medical issues such as gastrointestinal disorders, seizures, and sleep disturbances [2]. ASD affects 1 in 44 children by the age of 8 years in 2018 [3]. There have been recent concerns about alarming increases in the prevalence of autism spectrum disorder [4]. Such observed increase in ASD prevalence confirms that ASD is an urgent public health concern and emphasizes the need for continued surveillance based on consistent and reliable methods to monitor the rising prevalence of ASD.

There are many different ways to estimate the number of children with ASD. This estimate is in general referred to as prevalence, a scientific term that describes the number of people with a disease or condition among a defined group [5]. Prevalence is typically shown as a proportion (e.g., 1 in 1,000). The Centers for Disease Control and Prevention (CDC) began tracking and monitoring the prevalence of ASD in 1996, initially conducting studies among children in metropolitan Atlanta, Georgia [6]. Currently, four data sources related to ASD prevalence are presented on the Centers for Disease Control and Prevention (CDC) Autism Data Visualization Tool [5], including Special Education Child Count (SECC), National Survey of Children's Health (NSCH), Medicaid, and Autism and Developmental Disabilities Monitoring Network (ADDM). To the best of my knowledge, no research has been conducted to directly compare these four most common measures of autism prevalence. As a result, it is unclear if these four measures could generate consistent diagnostic data and if a combined use of these measures could improve ASD diagnosis. Effort needs to be taken to better understand the prevalence of autism spectrum disorders, allocate resources accordingly, and design our educational and healthcare systems appropriately. Furthermore, rationalization

of possible differences among different measures could improve ASD detection methods, which would further help to understand the causes of autism and develop possible preventive strategies.

## 2. Methods

The research methodology used in this study was a needs assessment. One of the study goals was to quantitatively assess the level of discrepancies among the four commonly used CDC measures of autism prevalence. Analysis was focused on data availability and coverage with respect to state and year. Afterwards, data collection methodology and criteria for each of the four measures of ASD prevalence were assessed. Based on the advantages and disadvantages of each measure, recommendations on potential improvement areas were proposed.

The raw data used in this study were taken directly from the CDC Autism Data Visualization Tool database [5], a federal source that houses reported ASD prevalence over time (between year of 2000 and 2020) for each US state and US total with respect to four different measures. The data were extracted and stored in a master Microsoft Excel file, sorted based on different ASD measures, then further subdivided by the year and the state. The data points were analyzed in detail, including computing the range, average, and standard deviation (or variance).

One of the most popular statistical analysis tools, T-Test, was also used to evaluate if any pair of ASD measures generated significantly different data. The T-Test is a type of inferential statistical analysis used to determine if there is a significant difference between two groups [7]. The following formulas were used to calculate T-value and Degrees of Freedom for an unequal variance T-Test [8]:

$$\text{T-value} = \frac{mean1 - mean2}{\sqrt{\left(\frac{var1}{n1} + \frac{var2}{n2}\right)}} \tag{1}$$

$$\text{Degrees of Freedom} = \frac{\left(\frac{var1^2}{n1} + \frac{var2^2}{n2}\right)^2}{\frac{\left(\frac{var1^2}{n1}\right)^2}{n1-1} + \frac{\left(\frac{var2^2}{n2}\right)^2}{n2-1}} \tag{2}$$

where:

mean1 and mean2 denote average values of sample set 1 and sample set 2, respectively.
var1 and var2 denote variances of sample set 1 and sample set 2, respectively.
n1 and n2 denote numbers of records in sample set 1 and sample set 2, respectively.

The critical value was determined by the widely available T-distribution table based on the degree of freedom value and the predetermined level of significance (i.e., p value, typical value of 5% is used). Afterwards, the critical value was compared with the T-value. If the T-value was bigger than the critical value, it was concluded that the two population sample sets had intrinsic differences that were statistically significant.

## 3. Results and Discussion

Based on the raw data extracted from the CDC database, numbers of available data points from four sources between 2000 and 2020 with respect to each state were summarized. In general, the ASD prevalence data from open sources were lacking, only tracing back 21 years. Data availability and the average number of yearly data were calculated and shown in Figure 1. Data availability was defined as the number of states/district (i.e., 50 states and Washington D.C.) where ASD data were available, while the number of yearly data was defined as the number of years between 2000 and 2020 when ASD data were available. It was clear that the SECC has the most available data points in general, with certain yearly data available from every state, mostly between 2000 and 2018. For the NSCH, while every state had some data, the available data over the year was very limited, as there were only 1 to 4 data points for each state (with average of 2). Medicaid had

the 2nd most available data points, with every state having some data - mostly 13 yearly data points between 2000 and 2012. The ADDM Network was only available for 19 out of the 51 states/district and the available data was limited on the year.

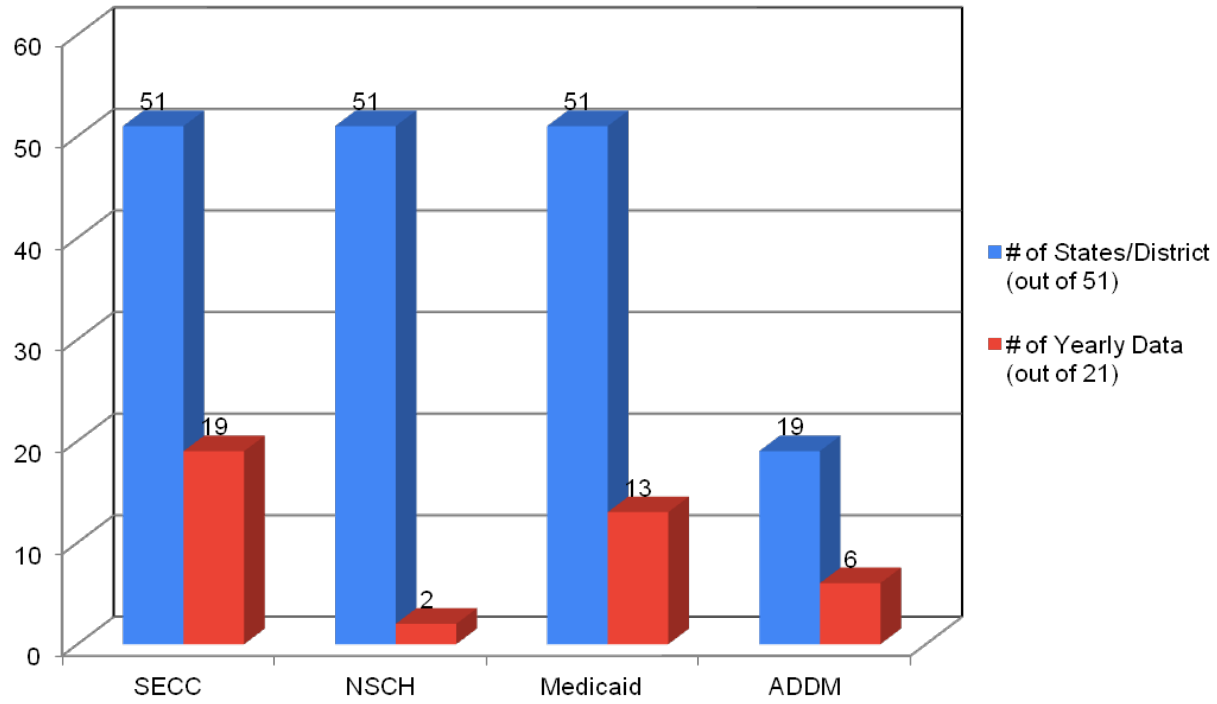


Figure 1. The data coverage in terms of number of states and average number of yearly data.

Figure 2 shows the ASD prevalence data over time for US total including plotted trendlines. The overall trend in the graph was very consistent across four data sources. In general, the reported prevalence of ASD had increased over the years, aligned with prior investigation by others [4]. Although the increase could be due to modifications to the clinical definition of ASD over time or better efforts to diagnose ASD in recent years, a true jump in the number of individuals with ASD was very possible, which was concerning to think about. So far, there is no clear evidence on what specifically was the dominant contributing factor. A more systematic long-term investigation might be required. From Figure 2, it was also evident that the data generated by different measures (SECC, NSCH, Medicaid, ADDM) showed some level of discrepancy. SECC and Medicaid were relatively similar. The ADDM network showed higher values compared to SECC and Medicaid, but the trend lines were quite similar. NSCH showed the most discrepancy; this combined with its limited data availability made NSCH a relatively unreliable data source.

The average ASD prevalence based on the different measures for each year was then calculated. A scatter plot (see Figure 3) displays the average value for ASD prevalence in each year and the error bars represent the standard deviation for each average. This diagram revealed that the standard deviations in the data were often sizable, confirming the discrepancy among different measures.

As part of further analysis, one year containing the most ASD prevalence data for all states related to four different measures was identified as the year of 2012 for more detailed investigation. Autism prevalence per 1000 children for each state related to four data sources in 2012 was summarized. Figure 4 shows the average values of ASD prevalence in 2012 with error bars representing standard deviations. This diagram revealed that the average values of these four different measures were different. However, it was hard to accurately judge if the four measures were significantly different or not because the standard deviations in the data were quite sizable. To properly address

this, statistical analysis on the 2012 data was performed by T-test (see Section Methods). As shown in Table 1, all pairs, except the SECC-Medicaid pair, demonstrated statistically significant differences.

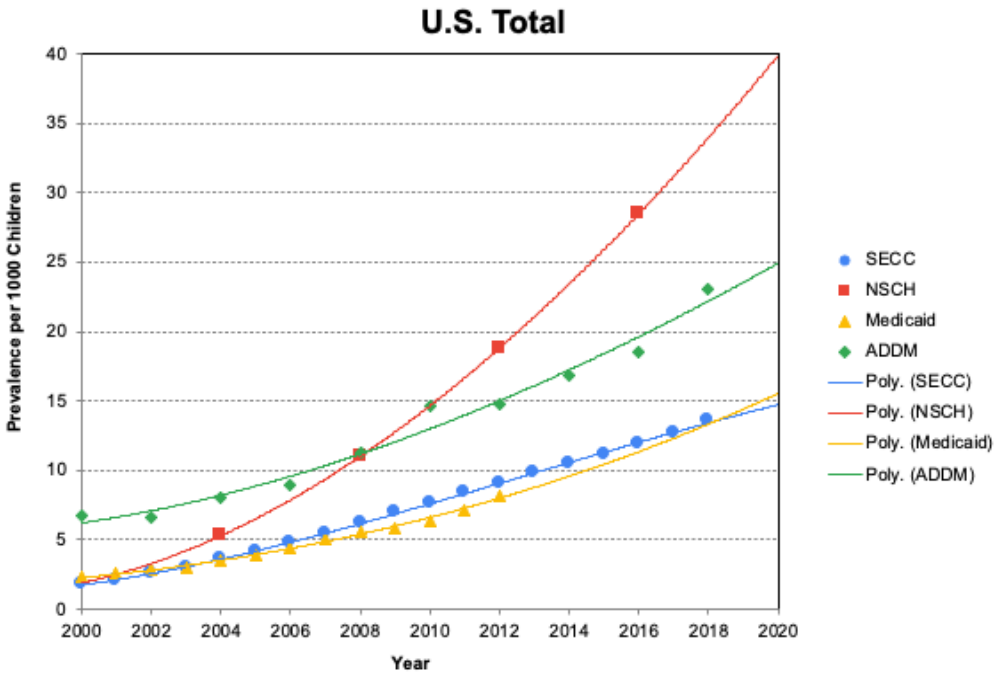


Figure 2. The ASD prevalence data over time for US Total. Lines represent the general trends.

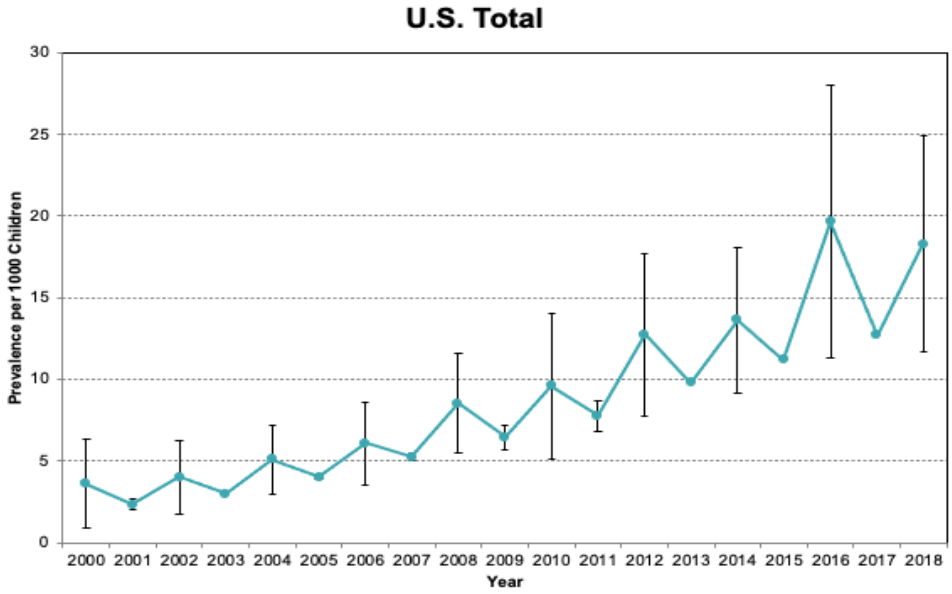


Figure 3. Average ASD prevalence based on different measures from 2000 to 2018 for US Total. Error bars represent the standard deviation of data.

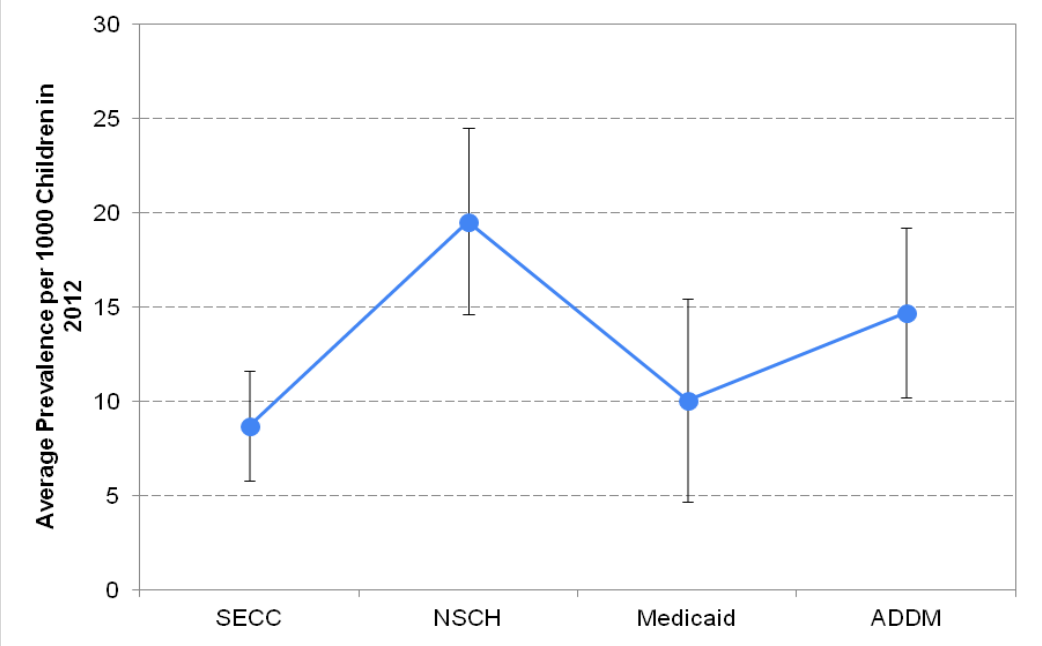

Figure 4. The average ADS prevalence with respect to different measures in 2012. Error bars represent the standard deviation of data.

Table 1. T-Test Calculation Result on different pairs of ASD prevalence measures (based on p value of 0.05).

| | **mean1** | **mean2** | **var1** | **var2** | **n1** | **n2** | **T-value** | **Degrees of Freedom** | **Critical value** | **Intrinsic difference** |
|---|---|---|---|---|---|---|---|---|---|---|
| **SECC vs NSCH** | 8.70 | 19.55 | 8.51 | 24.53 | 51 | 41 | 12.398 | 37.92 | 2.032 | yes |
| **SECC vs Medicaid** | 8.70 | 10.05 | 8.51 | 29.16 | 51 | 51 | 1.572 | 46.71 | 2.020 | no |
| **SECC vs ADDM** | 8.70 | 14.72 | 8.51 | 20.24 | 51 | 11 | 4.246 | 10.67 | 2.228 | yes |
| **NSCH vs Medicaid** | 19.55 | 10.05 | 24.53 | 29.16 | 41 | 51 | 8.777 | 47.71 | 2.018 | yes |
| **NSCH vs ADDM** | 19.55 | 14.72 | 24.53 | 20.24 | 41 | 11 | 3.094 | 17.54 | 2.110 | yes |
| **Medicaid vs ADDM** | 10.05 | 14.72 | 29.16 | 20.24 | 51 | 11 | 3.003 | 18.68 | 2.101 | yes |

Knowing that the ASD prevalence data based on different measures could be significantly different, the diagnosis design and data collection protocol related to each of these four data sources were investigated one-by-one with an

attempt to answer the following questions. Which data source was more representative and designed in a robust manner? How could these different measures be improved so that more consistent and reliable data were collected with minimized discrepancies?

The Special Education Child Count (SECC) was based on administrative data collected by the US Department of Education. The Individuals with Disabilities Education Act (IDEA) classified children of 3 - 21 years old with disabilities who received special education and related services into 13 primary disability categories, including ASD. CDC then used such data to report the number of children 6 - 17 years old with ASD who were receiving special education and related services in each state [5]. The data submitted by each state was reviewed and evaluated by the Office of Special Education Programs in terms of timeliness, completeness and accuracy [9]. Therefore, while the ADS prevalence data based on SECC could still be underreported, they should be relatively consistent and reliable, which was confirmed by the SECC data for several randomly chosen states (Arizona, Georgia & New Jersey). Under Section 618 of IDEA, all states were required to report the number of students who received special education and related services under the primary disability category for ASD. Therefore, national and state-level data were available annually for years 2000 - 2018, as validated by the earlier analysis (Figure 1). Surprisingly, it was noted that no SECC data post 2018 was recorded in the CDC database. Sufficient funding and resources should be secured so that the CDC database gets updated in a timely manner.

The National Survey of Children's Health (NSCH) is an annual, cross-sectional, address-based survey that collects information on the health and well-being of children ages 0-17 years [5]. The NSCH is funded and directed by the Health Resources and Services Administration's Maternal and Child Health Bureau and fielded by the US Census Bureau [10]. The data were collected via telephone for early survey years (2003, 2007, and 2011-12) but have been based on both web-based and paper and pencil methodologies since the beginning of 2016 [11]. In general, conducting surveys was still one of the most straightforward ways to gather useful information if a robust design (including sufficient population and time coverage) was implemented. However, based on earlier findings (Table A and Figure 1), while NSCH covered all US states, every state has only data available for 1 to 3 years between 2000 and 2020. The data coverage and frequency of survey needed to be significantly improved so that the data collection based on NSCH can be re-evaluated in the future.

The ASD prevalence from Medicaid was based on administrative claims data from the Centers for Medicare and Medicaid Services (CMS). The CDC analyzed the Medicaid Analytic eXtract (MAX) datasets released by CMS for the years 2000 - 2012, and identified children 3 - 17 years old who have received Medicaid benefits and had at least two outpatient billing codes for ASD or one inpatient billing code in the specified year [5]. Figure 5 shows the ASD prevalence data based on Medicaid data source for the States of Mississippi, New Hampshire and Maine. There was a sizable difference between different states, likely due to the fact that the measurement of poverty varied state to state. However, the ASD prevalence was not positively correlated to the poverty rate of each state. Even though the state of Mississippi had the highest poverty rate of 19.78% while the state of New Hampshire had the lowest of 7.42% [12], the ASD prevalence based on Medicaid for Mississippi was much lower than New Hampshire. It is likely that states with higher poverty rates could establish stricter income eligibility for Medicaid although more thorough data analysis is required to validate such hypotheses. Because this method was so heavily dependent on income eligibility set by each state, the design of collecting data and criteria threshold should be fully aligned among different states.

The Autism and Developmental Disabilities Monitoring (ADDM) Network is an active surveillance system that provides estimates of the prevalence of autism spectrum disorder among children aged 8 years whose parents or guardians reside within ADDM sites in the United States [4]. While the data generated had some fluctuations over the years, it is still believed that the ADDM Network is the most robust ASD surveillance system in the United States among the four commonly used measures. It provided not only the ASD prevalence estimates for specific areas, but also information related to geographical variations and subgroup breakdowns defined by sex and race/ethnicity. Furthermore, it was the only data collection method to incorporate ASD diagnostic criteria into the case definition rather than relying completely on the reports from parents and caregivers. Since the ADDM Network was directly funded by CDC, the data collection protocol and future activities could be constantly reviewed and enhanced based on the prior learning. Currently, there are only 19 states containing ADDM Network Surveillance Sites, and the data are only collected every other year. CDC should secure

sufficient funding to extend the coverage to include all states and capture ASD prevalence data every year considering the importance of such a measure to the ASD community.

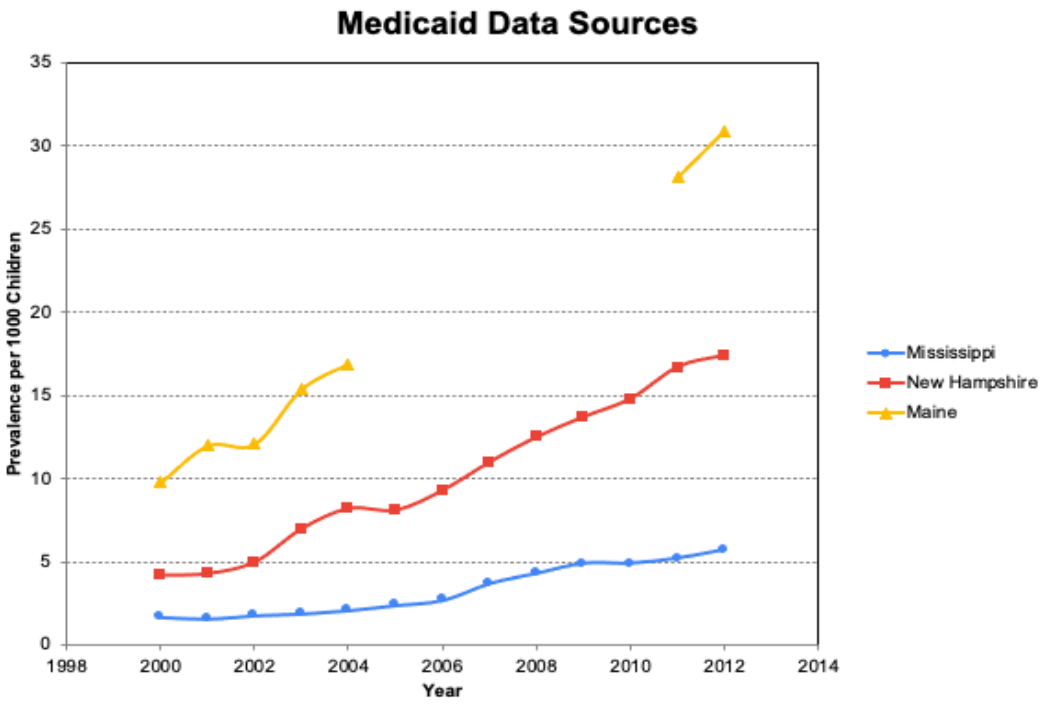


Figure 5. ASD prevalence data based on Medicaid measure for States of Mississippi, New Hampshire and Maine.

## 4. Conclusion

In summary, this study revealed that, in general, the ASD prevalence data based on publicly available data sources were far from being sufficient. The four commonly used ASD data sources (i.e., Special Education Child Count, National Survey of Children's Health, Medicaid, Autism and Developmental Disabilities Monitoring Network) did not cover all states within the United States and all the years between 2000 and 2020. This study also clearly demonstrated that there was a significant disparity among the four commonly used measures, although they were the only data sources being publicly used. Under such conclusions, there is a strong desire for additional and more prompt efforts to increase the data coverage and reduce the disparities through better policy alignment and methodology improvement.

Among those four commonly used measures of ASD prevalence, the ADDM Network was the most robust data source based on the overall design while the SECC seemed to be the most consistent and liable. All four measures had their specific pros and cons. Based on this study, the proposed recommendations to improve the four measures were presented in Table 2. In general, more funding and resources were strongly urged, with special attention paid to better population and time coverage, research to lower initial age of autism diagnosis, and methodology improvement related to the SECC and ADDM data sources.

Moreover, policy makers and relevant government authorities need to encourage uniformity in all ASD measurement requirements across the states. Regular updating of federal policies is needed to solidify the alignment with current measurement methods to encourage and allow for earlier and more accurate identification of children with ASD.

Table 2. The pros and cons, and recommended improvement areas for four data sources related to ASD prevalence.

| Data Sources | SECC | NSCH | Medicaid | ADDM Network |
|---|---|---|---|---|
| **Funded and Administered by** | US Department of Education | Health Resources and Services Administration's Maternal and Child Health Bureau | Centers for Medicare and Medicaid Services (CMS) | Centers for Disease Control and Prevention (CDC) |
| **Target Children** | 6 - 17 years old | 0 - 17 years old | 3 -17 years old | 8 years old |
| **Pros** | Data seems consistent and reliable;<br>States are required to report the number | Survey is still the most straightforward way to generate useful information in general | Relatively better data availability | Scientific way to obtain data based on method design |
| **Cons** | Data are available only for years 2000 - 2018 | With such limited data, it is hard to judge the robustness of data collection. | Medicaid is for those in poverty, measurement of poverty (income eligibility) varies state to state | ADDM is missing data for a lot of states; data available only every other year |
| **Improvement Recommendation** | Make sure sufficient funding is allocated for yearly recording and timely update | The coverage and frequency of survey needs to be improved so that this measure can be re-evaluated in the future | The design of collecting data may need to be revisited and improved | Data seems consistent and reliable;<br>More funding is required for all states to participate |

## References


[1] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders, 5th Ed*. Arlington, VA: American Psychiatric Publishing, 2013.

[2] H. Hodges, C. Fealko, N. Soares, "Autism Spectrum Disorder: Definition, Epidemiology, Causes, and Clinical Evaluation", *Translational Pediatrics*, vol. 9, suppl 1, pp. S55-65, 2020.

[3] M. J. Maenner, et al, "Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016". *Morbidity and Mortality Weekly Report*, vol. 69, no. 4, pp. 1-12, 2020.

[4] J. Baio, et al, "Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014", *MMWR Surveill Summ.*, vol. 67, no. 6, pp. 1-23, 2018.

[5] CDC (2021), Autism Spectrum Disorder (ASD) [Online]: https://www.cdc.gov/ncbddd/autism/data/index.html.

[6] M. Yeargin-Allsopp, et al, "Prevalence of Autism in a US Metropolitan Area". *JAMA*, vol. 289, pp. 49–55, 2003.

[7] A. Ross and V. L. Willson, *Basic and Advanced Statistical Tests*, Rotterdam: SensePublishers, 2017.

[8] A. Hayes (2022), "T-Test", Investopedia [Online]: https://www.investopedia.com/terms/t/t-test.asp. 2022.

[9] OSEP (2020), IDEA Part B Child Count and Educational Environments for School Year 2019-2020 [Online]: https://www2.ed.gov/programs/osepidea/618-data/collection-documentation/data-documentation-files/part-b/child-count-and-educational-environment/idea-partb-childcountandedenvironment-2019-20.pdf.

[10] HRSA (2022), National Survey of Children's Health (NSCH) [Online]: https://mchb.hrsa.gov/data-research/national-survey-childrens-health. 2022.

[11] R. M. Ghandour, et al, "The Design and Implementation of the 2016 National Survey of children's Health", *Matern child Health J.* vol. 22, no. 8, pp. 1093-1102, 2018.

[12] World Population Review (2022), Poverty Rate by State 2022 [Online]: https://worldpopulationreview.com/state-rankings/poverty-rate-by-state.