

# Evaluation of the Completeness of ALS Case Ascertainment in the US National ALS Registry: Application of the Capture-Recapture Method

Lorene M. Nelson<sup>a</sup> Barbara Topol<sup>a</sup> Wendy Kaye<sup>b</sup> Jaime Raymond<sup>c</sup>  
D. Kevin Horton<sup>c</sup> Paul Mehta<sup>c</sup> Todd Wagner<sup>d, e</sup>

<sup>a</sup>Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA, USA; <sup>b</sup>McKING Consulting Corporation, Atlanta, GA, USA; <sup>c</sup>Division of Toxicology and Human Health Sciences, Agency for Toxic Substances and Disease Registry/Centers for Disease Control and Prevention, Atlanta, GA, USA; <sup>d</sup>Department of Surgery, Stanford University School of Medicine, Stanford University School of Medicine, Stanford, CA, USA; <sup>e</sup>Health Economics Resource Center, Veterans Affairs Palo Alto Health Care System, Menlo Park, CA, USA

## Keywords

Amyotrophic lateral sclerosis · Motor neuron disease · Prevalence · Capture-recapture method · Administrative data

## Abstract

**Introduction:** The Centers for Disease Control and Prevention (CDC) National Amyotrophic Lateral Sclerosis (ALS) Registry is the first national registry for a chronic neurologic disease in the USA and uses a combination of case-finding methods including administrative healthcare data and patient self-registration. **Methods:** We applied capture-recapture methodology to estimate the completeness of the Registry for ascertaining patients with ALS for the first full year and the fourth year of the Registry (2011, 2014). The Registry uses the combination of two national administrative claims databases (Medicare and Veterans Affairs) with a self-register option at the registry portal. We conducted descriptive analyses of the demographic and clinical characteristics of the ALS cases identified by each of the sources and estimated the completeness of case ascertainment for each of the

three ALS Registry sources individually, pairwise, and in all combinations. **Results:** Case-finding completeness was 54% in 2011 and improved to 56% in 2014. A smaller proportion of ALS patients under age 65 were ascertained than those 65 or older, and ascertainment was also lower for nonwhite than white patients. The uncorrected ALS prevalence was 4.3/100,000 in 2011 (in 2014, 5.0/100,000), but after correction for underascertainment, annual prevalence in 2011 was 7.9/100,000 (95% CI: 7.6–8.2) (in 2014 was 8.9/100,000 [95% CI: 8.7–9.2]). **Discussion/Conclusion:** Our findings indicate that administrative healthcare databases are a very efficient method for identifying the majority of ALS prevalent cases in the National ALS Registry and that the inclusion of a web registry portal for patients to self-register is important to ensure a more representative population for estimating ALS prevalence. Nonetheless, more than 40% of ALS cases were not ascertained by the Registry, with individuals younger than age 65 and people of color underrepresented. Recommendations are provided for additional methods that can be considered to improve the completeness of case ascertainment.

© 2021 The Author(s)  
Published by S. Karger AG, Basel

## Introduction

Despite the importance of amyotrophic lateral sclerosis (ALS) as the third most common neurodegenerative disorder of aging, efforts to estimate the prevalence of ALS in the USA have been challenging, and no reliable national estimates have been available before 2010. The number of people with ALS in the USA is difficult to estimate because ALS, like most noncommunicable diseases, is not a notifiable condition. Furthermore, until recently, there was no nationwide disease registry or surveillance system for ALS. Prior estimates of national ALS prevalence were obtained by extrapolating findings from prevalence studies done in smaller geographic regions to the nation as a whole. While estimates of national ALS prevalence have been possible in other countries that have universal healthcare systems [1–5], to date, it has been challenging to obtain an estimate of ALS prevalence in the USA.

In recent years, there has been increasing interest in establishing national surveillance efforts to estimate the national prevalence of neurologic disease. These efforts began with the establishment of the National ALS Registry, which was mandated by the US Congress in 2008, for the purpose of estimating the incidence and prevalence of ALS, characterizing the demographics of those living with the disease, and examining the risk factors for ALS [6]. More recently, with the 21st Century Cures Act of 2016 [7], the Senate authorized the Centers for Disease Control and Prevention (CDC) to create a similar surveillance system for other neurologic diseases beginning with multiple sclerosis and Parkinson's disease, making it important to examine the most efficient and valid methods for identifying persons with chronic neurologic disease in the USA.

CDC's Agency for Toxic Substances and Disease Registry (ATSDR) began methodologic work to establish the National ALS Registry in 2009 and sought the best methods for identifying patients with ALS. Through early pilot efforts, the Registry developed a novel case-finding method to identify patients with ALS using the combination of several national healthcare administrative databases, supplemented with a mechanism by which ALS patients could self-register at the Registry's national portal [8]. The self-report option was designed to improve capture for people with private health insurance, given that ALS is frequently diagnosed between the ages of 40 and 70 when people are still employed. With the goal of evaluating the completeness of case ascertainment using these methods, the Registry conducted a series of independent

studies between 2009 and 2011 in eight US states and three metropolitan regions [9–14]. A subsequent report by the Registry determined that 54% of ALS patients identified by active case-finding in those studies were able to be ascertained using registry methodology [15].

Capture-recapture is a statistical method that examines the overlap in identification of cases from different data sources and uses this overlap to estimate disease prevalence [16, 17]. In epidemiology, capture-recapture methodology is widely used as a method to estimate the number of cases who were not identified by any of the case-finding methods, thus enabling a conclusion about the completeness of ascertainment. This method is often the only method available other than to carry out very costly and time-consuming validation studies where all cases in a smaller geographic region can be identified through contact and participation of all neurologic care providers.

We had three primary objectives: (i) to use capture-recapture modeling to estimate the degree of case ascertainment completeness for the first and fourth year of the Registry (2011, 2014) to obtain two separate estimates of completeness with the latter one reflecting a mature Registry, (ii) to determine the relative effectiveness of the Registry's case-finding methods for identifying persons with ALS and to determine if any groups were systematically underrepresented, and (iii) to obtain revised estimates of ALS prevalence for subsequent years after correcting for underascertainment. We sought to compare the statistical estimation of case capture completeness using capture-recapture methodology with the estimates obtained from independent studies conducted in state and metropolitan regions. Going forward, these estimates will be important for obtaining ascertainment-adjusted annual estimates of ALS prevalence for the Registry and for informing future neurological surveillance systems in the USA.

## Methods

We used two national administrative databases to estimate ALS prevalence: Medicare (parts A and B) and Veterans Affairs data (comprising Veterans Health Administration [VHA] and Veterans Benefit Administration [VBA] data). Because the majority of VBA patients had been seen in the VHA for medical care, we combined VHA and VBA into a single Veterans Affairs (VA) source for evaluating the completeness of case ascertainment. Approximately 11% of patients seen by the VA are nonveterans who qualify for health care coverage through the VA Civilian Health and Medical Program (CHAMP-VA) [18], many of whom may have commercial insurance and/or qualify for Medicare.

The administrative data files include computerized records documenting outpatient visits, inpatient hospitalizations, clinical procedures, laboratory results, and medication prescriptions. Potential ALS cases were identified by actively searching the databases for ICD-9 (International Classification of Diseases, 9th Revision, Clinical Modification) codes of 335.2X (motor neuron diseases [MND]), ALS [ICD 335.20], progressive muscular atrophy (PMA; ICD-9 code 335.21), progressive bulbar palsy (PBP; ICD-9 code 335.22), pseudobulbar palsy (PBP; ICD-9 code 335.23), primary lateral sclerosis (PLS; ICD-9 code 335.24), and other adult-onset motor neuron diseases (other MND; ICD-9 code 335.29). We also obtained information from pharmacy data (available in the Medicare and VA sources) indicating whether potential cases had been prescribed riluzole (brand name Rilutek), the only FDA-approved medication indicated for ALS during this period. We considered using Medicaid to identify ALS patients; however, an earlier study by our group for the period 2002–2004 found that when Medicare and the VA are used as case-finding methods, the addition of Medicaid uniquely identified only 3% of cases [19].

The ALS case identification algorithm was developed from work done through several pilot projects [8]. Patients were categorized into three distinct groups: definite ALS, possible ALS, and not ALS. The algorithm was developed by using several different criteria in an iterative process to determine the best model for categorizing patients. It was possible for a given patient to satisfy more than one criterion to be classified as definite ALS (online suppl. Appendix 2 Table A6; for all online suppl. material, see [www.karger.com/doi/10.1159/000521591](http://www.karger.com/doi/10.1159/000521591)). Six criteria were used for classifying patients as definite ALS, and all patients satisfied at least one of the six criteria: criterion 1: having an ICD-9 code of ALS in one or more years from the same case-finding source and/or prescription for riluzole; criterion 2: an ALS code in two or more years with a neurologist visit in the same source; criterion 3: age <65, with one or more ALS codes in Medicare and  $\geq 1$  neurology visit; criterion 4, one or more ALS codes in  $\geq 1$  year with at least one neurologist visit in one source, combined with 1 or more ALS codes in another source; criterion 5: having an ALS code in 3 or more sources; and criterion 6: having an ALS code in one year and 5 or more neurologist visits. The validated algorithm has been shown to have high sensitivity (87%) and specificity (85%) [8].

Deduplication of subjects contained in more than one administrative data source was carried out at ATSDR based on identifiers such as name, date of birth, sex, and the last 5 digits of the social security number. We used the personal identifiers to match subjects between data sources in order to determine the unique combination of sources that identified each ALS case. Also, because we had identifiers from all three data sources, if a demographic variable such as race was missing in one source, we were able to obtain that information for individuals who were present in one or both of the other sources. As a result, in the analytic data set, there were very little missing data for sex (1.6%) and age (0.3%) and 8.9% missing information for race.

The Registry was previously approved by CDC/ATSDR's IRB for obtaining identified data from all sources and for carrying out the matching process that required personal identifiers. CDC analysts prepared an analytic data set containing deidentified information for the Stanford University study team and included information on demographic characteristics and case definition criteria for application of the capture-recapture method. We received an exemption from the Stanford University Institutional

Review Board (IRB) approval because deidentified information data were utilized and the re-constitution of patients' identities was not possible. We carried out statistical analyses using SAS<sup>TM</sup> software [20].

#### *Application of the Capture-Recapture Method*

The history, mathematical formulation, and assumptions underlying the capture-recapture method are the subject of an excellent review [21] and are described in detail in online supplementary Appendix Supplement 1. In brief, we conducted descriptive analyses of the demographic and clinical characteristics of the ALS cases identified by each of the sources for each of the three Registry sources (Medicare, VA, and registry portal), pairwise, and in various combinations. We characterized the degree of dependence between each pairwise combination of sources by multiplying the proportion of cases identified by each source to estimate the proportion that would be expected if the two sources were operating independently. We then compared "expected proportion under independence" to the observed proportion of cases identified by both sources. Positive dependency occurs when the probability of being "captured" or identified by one source is positively associated with the probability of being identified by the other source. Negative dependency occurs when the probability of being identified by one source is inversely associated with the probability of being identified by the other source.

Because it is unlikely that any two sources provide a completely independent capture of cases, we carried out log-linear modeling [22] using the SAS<sup>TM</sup> GENMOD procedure to calculate capture-recapture estimates derived by various combinations of the three sources. Our base model included a regression term for each source to represent its main effect (*model 1*), and we modeled source dependence between any two sources by singly adding the corresponding two-source interaction term to the model with the main effects (*models 2–4*). Then, we fitted models with main effects and two of the three two-source interactions terms (*models 5–7*), and lastly, we estimated the number of people with ALS who were not identified by any of the three sources by fitting a fully saturated model including all two-source and three-source interactions (*model 8*). The significance of each interaction term was assessed using the likelihood ratio statistic ( $G^2$ ). Goodness of fit was assessed by examining the deviance of alternative models, and the best model was selected by choosing the one with the smallest Akaike Information Criterion (AIC). When two or more models had a similar degree of fit, we chose the one with the fewest interaction terms (i.e., most parsimonious) as the final model. Confidence intervals (CIs) for estimated number of cases were computed using the profile likelihood method [23].

The strictest application of capture-recapture assumes a homogenous study population where all members have the same probability of being included in each source. In practice, however, ascertainment can vary according to factors such as sex, age, and race. Therefore, we applied capture-recapture methodology in separate demographic groups to adjust for the number of unobserved patients according to sex, age group, race, and year of study. In each stratum-specific model, we identified the minimum AIC model and chose the model with the lowest AIC. We fit log-linear models to estimate the number of missing cases that allowed for dependencies among the sources and for effects of sex, age, and race on the probability of an individual being identified from a source.

**Table 1.** Demographic characteristics of ALS cases according to the case ascertainment source, ALS prevalent cases, 2014

Characteristic	Medicare		Veterans Affairs (VHA, VBA)		ALS Registry portal		Total of all sources	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Cases identified by source, <i>n</i>	11,898	74.7	3,586	22.5	5,317	33.4	15,927	100
Sex								
Male	7,026	60.2	3,390	96.7	3,216	60.5	9,821	62.7
Female	4,639	39.8	117	3.3	2,101	39.5	5,854	37.3
Missing	(233)	(2.0)	(19)	(0.5)	(0)	(0.0)	(252)	(1.6)
Age								
Median/mean age	67.0/65.7		68.0/67.5		61.0/60.4		65.0/64.2	
<65	4,972	41.8	1,162	32.5	3,318	62.8	7,643	48.1
≥65	6,926	58.2	2,410	67.5	1,962	37.2	8,233	51.9
Missing	(0)	(0)	(14)	(0.4)	(37)	(0.7)	(51)	(0.3)
Race								
White	10,212	86.7	3,104	89.3	3,753	91.9	12,660	87.2
Nonwhite	1,569	13.3	370	10.7	331	8.1	1,851	12.8
Missing	(117)	(1.0)	(112)	(3.1)	(1,233)	(23.2)	(1,416)	(8.9)

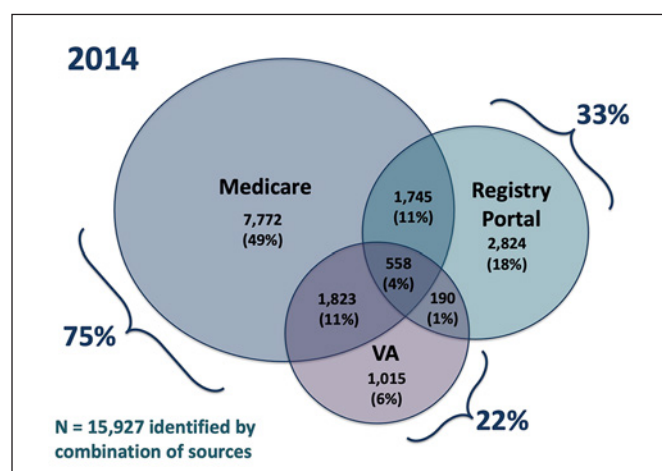
We calculated ascertainment-corrected point prevalence estimates of ALS for the years 2011 and 2014, using capture-recapture methods to correct for underascertainment in each stratum defined on the basis of age, sex, and race. We used the parameter estimates from capture-recapture log-linear models to estimate the total number of ALS cases in each population subgroup (defined on the basis of age, sex, and race).

## Results

The demographic characteristics of ALS cases identified in 2014 are presented in Table 1 (2011 results in online suppl. Table A2). Males comprised 95% of cases identified through the VA and 52% of the non-VA cases. In the registry portal source, only 37% of cases were aged 65 and above, whereas this group was considerably larger in Medicare (58%) and the VA (67%). In the total group, 12% of ALS cases were nonwhite. The percent of nonwhites was somewhat lower in the registry portal-identified cases (8%) than in Medicare (13%) or VA (11%).

The number of ALS cases identified by each source by itself, as well as by any two- or three-source combinations of sources in 2014, is summarized in Figure 1 (2011 results in online suppl. Fig. A1). In 2014, the highest yield source was Medicare (75% of all cases), followed by registry portal (33%) and VA (23%). The majority of cases (73%) were identified by only one source.

When two sources are negatively dependent, the probability of being captured by one source is less likely given



**Fig. 1.** Number and percentage of ALS prevalent cases identified by combinations of sources (National ALS Registry, 2014).

the case was captured by the second source. Examination of Table 2 indicates that when the case-finding sources were examined pairwise, they were all negatively dependent. The strongest negative dependency is between VA and Registry portal because the expected probability (underindependence) of being identified by both was 0.287, considerably higher than the observed proportion identified by both (0.09). This strong difference between the observed proportion and the expected probability indicates that cases identified by the portal are not likely to be identified by VA, and vice versa.

**Table 2.** Examination of two-source dependencies, ALS prevalent cases, 2014

Two-source combination	Number in category	Probability of being identified by first source	Probability of being identified by second source	Expected probability of being identified by both sources (assuming independence)	Observed proportion identified by both sources	Observed minus expected probability of being in both sources	Direction of dependence
Medicare and VA	13,103	0.908	0.273	0.248	0.182	-0.066	Negative
Medicare and portal	14,912	0.798	0.357	0.285	0.154	-0.130	Negative
VA and portal	8,155	0.652	0.440	0.287	0.092	-0.195	Negative

**Table 3.** Three-source capture-recapture estimates using log-linear modeling, ALS prevalent cases, 2014

Model	Terms	df	AIC	Estimated <sup>a</sup> , <i>n</i>	95% CI	Estimated <sup>b</sup> , <i>N</i>	95% CI
<b>Main effects only</b>							
1	Medicare, portal, VA	3	756.6	7,753	7,403–8,119	23,680	23,330–24,046
<b>Main effects with 2-source interactions one-at-a-time</b>							
2	Medicare, portal, VA, Medicare*portal	2	484.1	4,872	4,521–5,250	20,799	20,448–21,177
3	Medicare, portal, VA, Medicare*VA	2	113.4	12,019	11,351–12,725	27,946	27,278–28,652
4	Medicare, portal, VA, portal*VA	2	751.7	7,589	7,228–7,969	23,516	23,155–23,896
<b>Main effects with 2-source interactions</b>							
5	Medicare, portal, VA, Medicare*portal, Medicare*VA	1	105.8	15,086	12,865–17,691	31,013	28,792–33,618
6	Medicare, portal, VA, Medicare*portal, portal*VA	1	426.3	4,327	3,995–4,687	20,254	19,922–20,614
7	Medicare, portal, VA, Medicare*VA, portal*VA	1	104.8	12,578	11,802–13,405	28,505	27,729–29,332
<b>Saturated model</b>							
8	Medicare, portal, VA, Medicare*portal, Medicare*VA, portal*VA	0	77.2	20,567	16,966–24,932	36,494	32,893–40,859

AIC, Akaike Information Criterion. <sup>a</sup> *n* is the estimated unobserved cell using log-linear modeling, i.e., the estimated number not identified by any of the 3 sources. <sup>b</sup> *N* is the estimated total population using log-linear modeling, i.e., the total number of cases actually identified by the 3 sources plus the estimated missing cell *n*.

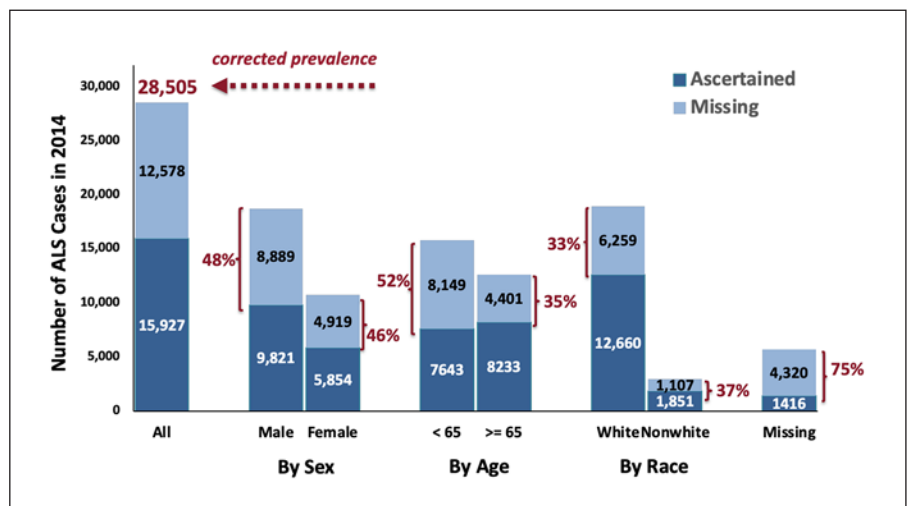
We carried out log-linear modeling to calculate capture-recapture estimates derived by combinations of the three sources (Medicare, VA, and registry portal) for the year 2014 (Table 3) (online suppl. Table A3 for 2011 results). In Table 3, “*n*” is the estimated unobserved cell using log-linear modeling, i.e., the estimated number of ALS cases not identified by any of the three sources. “*N*” is the estimated total population using log-linear modeling (i.e., the total number of cases actually identified by the three sources plus the estimated missing cell *n*). For both years of the registry, the best fitting model was model 7 (i.e., the one with the smallest AIC), which contained terms for the three main effects (Medicare, registry portal, and VA) and two of the two-source interaction terms (Medicare\*VA, portal\*VA). The inclusion of these interaction terms in model 7 adjusted for the negative dependencies between the 2 source pairs: (Medicare\*VA) and (registry portal\*VA). Notably, very similar results were obtained with log-linear models for the years 2011 and 2014.

Estimates of the percentage of case completeness were similar for 2011 and 2014 (54.1% and 55.9%, respectively). In 2011, the estimate of the missing cell count (i.e., the count of those not identified by any of the three sources) was 11,268, indicating that 45.9% of all ALS cases (11,268 of the corrected total 24,550 cases) were missed by all three methods (online suppl. Table 5a). In 2014, the estimate of the missing cell count was 12,578, indicating that 44.1% of all ALS cases (12,578 of the corrected total 28,505 cases) were missed by all three case-finding methods (online suppl. Table 5b). Thus, the percentage missing decreased by approximately 2% between the first and fourth years of the registry, suggesting that the estimate of underascertainment can continue to be used for several more years since the same case-finding methods are in place. Figure 2 presents the annual number of ALS cases in 2010–11 and 2014 by sex (and total), corrected for case undercount using the capture-recapture method.

**Table 4.** Observed count, missing cell estimate, percent missing, and uncorrected and corrected ALS prevalence, 2014

Characteristic	Observed count	Missing cell estimate*	Corrected population estimate	Missed, %	Population denominator	Uncorrected prevalence estimate			Corrected prevalence estimate		
						prevalence	lower 95% CI	upper 95% CI	prevalence	lower 95% CI	upper 95% CI
All	15,927	12,578	28,505	44.1	318,857,056	5.0	4.9	5.1	8.9	8.7	9.2
Sex											
Male	9,821	8,889	18,710	47.5	156,877,672	6.3	6.1	6.4	11.9	11.1	13.0
Female	5,854	4,919	10,773	45.7	161,979,384	3.6	3.5	3.7	6.7	6.4	6.9
Missing	252	25.0	277	8.9							
Age											
<65 years	7,643	8,149	15,792	51.6	272,642,163	2.8	2.7	2.9	5.8	5.5	6.1
≥65 years	8,233	4,401	12,634	34.8	46,214,893	17.8	17.4	18.2	27.3	25.3	29.9
Missing	51	10	61	16.3							
Race											
White	12,660	6,259	18,919	33.1	233,963,128	5.4	5.3	5.5	8.1	7.9	8.3
Nonwhite	1,851	1,107	2,958	37.4	84,893,928	2.2	2.1	2.3	3.5	3.2	3.8
Missing	1,416	4,320	5,736	75.3							

\* The estimated missing cell estimate, the number of cases not identified by any of the 3 sources, was obtained using log-linear modeling. All missing cell estimates were obtained from model 7, the model that consistently had the lowest AIC across strata.



**Fig. 2.** Number of ALS cases in 2014 by sex, age, and race, corrected for case undercount using the capture-recapture method. ALS, amyotrophic lateral sclerosis.

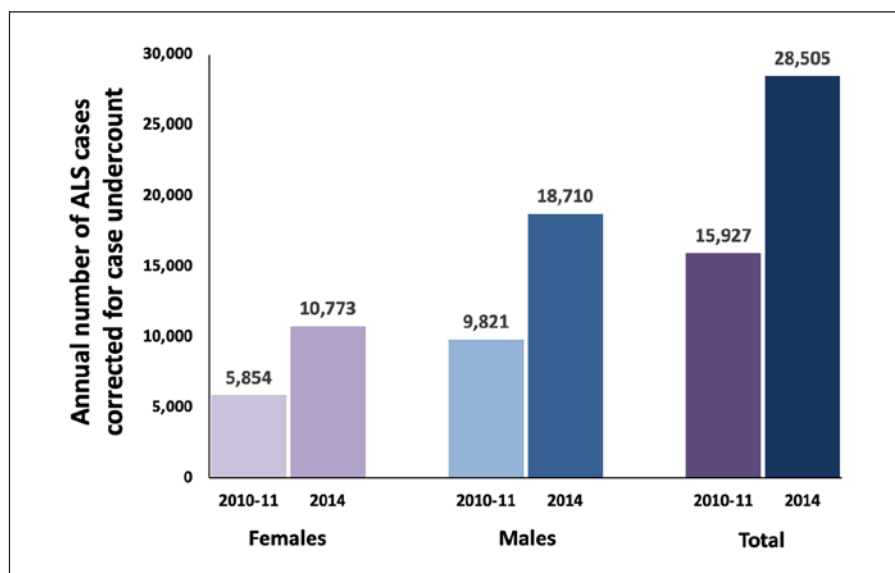
Table 4 and Figure 3 contain estimates of the unobserved percentage (missing %) by the demographic subgroup (and in online suppl. Appendix 2 Table A4 for 2011), indicating that the unobserved percentage was comparable for men and women; however, case completeness varied importantly according to age and race. The unobserved percentage was lower for patients aged 65 and above (35%) than it was for patients under age 65 (52%). It was slightly lower for whites (33%) than nonwhites (37%), and the lowest degree of case completeness

was estimated for cases whose race was not recorded in the administrative healthcare databases (75% for those who had no recorded race).

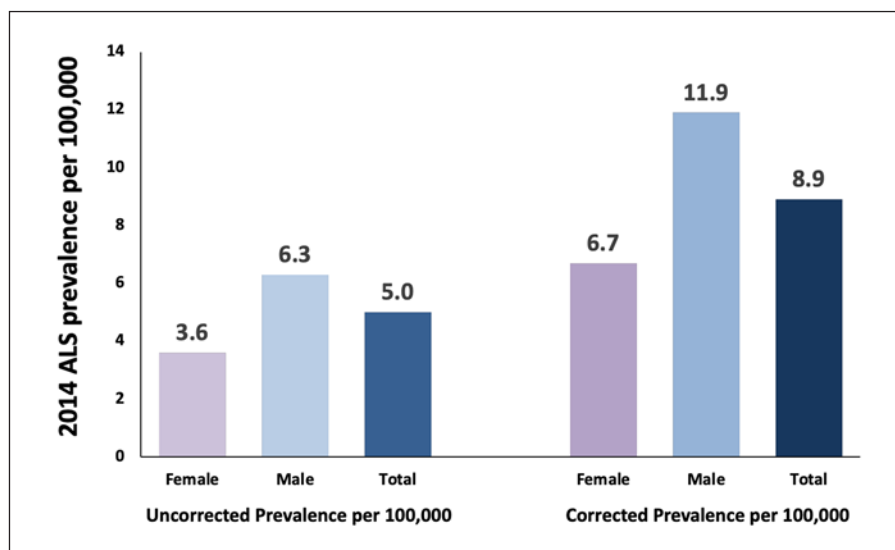
#### Corrected Prevalence Estimates

The uncorrected ALS prevalence estimates for 2010–11 and 2014 were 4.3 per 100,000 (95% CI: 4.2–4.3; online suppl. Appendix 2 Table A4) and 5.0/100,000 (95% CI: 4.9–5.1; Table 4; Fig. 4), respectively. After correction for the counts of unobserved cases from the capture-recap-

**Fig. 3.** Number of ALS cases, in 2010–11 and 2014, corrected for case ascertainment using the capture-recapture method. ALS, amyotrophic lateral sclerosis.



**Fig. 4.** Annual ALS prevalence estimates, 2014, uncorrected and corrected for case undercount using the capture-recapture method. ALS, amyotrophic lateral sclerosis.



ture model, the prevalence increased to 7.9 per 100,000 in 2010–11 (95% CI: 7.6–8.2) and to 8.9/100,000 (95% CI: 8.7–9.2) in 2014. In subsequent annual reports for the registry, prevalence estimates may be corrected for the degree of underascertainment estimated by our models (45%).

## Discussion

The US National Amyotrophic Lateral Sclerosis Registry uses the combination of two national administrative claims databases with a self-register option at the registry

portal. To our knowledge, no other registry prior to this one has been established to estimate the prevalence of a neurologic disease in the USA. With our application of capture-recapture methods, we estimated that the three case-finding methods together (Medicare, VA, and registry portal) ascertained approximately 54–56% of ALS cases in the USA and that the proportion of completeness did not differ between the first and fourth registry years. This estimate is virtually identical to the estimate of registry completeness that was obtained when Registry scientists used a different method to estimate the proportion of ALS cases that could be ascertained using case-finding methodology at the time the Registry was first established.

They conducted rigorous population-based studies in several state and metropolitan areas during the time period from October 2010 to December 2011, using intensive case-finding methods combined with case definition validation by examining patient's medical records to identify all possible ALS cases residing in the regions under study [15]. Of the total ALS cases documented in those population-based studies, 57% (2,720/4,767) were able to be matched to cases that had been ascertained using the registry case-finding methods. The fact that the estimates of case ascertainment completeness were nearly identical when using two very different methods of estimation lends credibility that our capture-recapture estimate that 54–56% of cases are ascertained by the Registry is robust.

One possible explanation for the observed undercount is the need to rely on an algorithm to identify patients with definite ALS in the administration databases. If the validated algorithm missed true definite ALS cases in any given year of the registry, then it would lead to an undercount. The degree of undercount attributed to the algorithm can be estimated by the number of cases who were classified as definite in any given year that were ascertained but identified as possible ALS in the previous year and correcting the previous year's estimate for that undercount. For example, in 2015, we identified 1,596 cases who were definite ALS who had been possible ALS in the year 2014. This indicates that 12.7% (1,596 out of 12,578) of estimated missing cases in 2014 could be attributed to the algorithm failing to identify the cases as definite ALS in 2014. If we add this undercount to the verified case count in 2014, we would have ascertained 17,523 of the capture-recapture estimated total 28,505 cases, or 61% of all cases.

Additional insights into case-finding methodology can be gained by examining how the completeness of case-finding varied by demographic subgroups. We examined how the completeness of ascertainment varied according to sex, age, and race and found that there were important differences in the case completeness of the sources. Although case-finding completeness did not differ much by sex, it was lower for people under age 65 years (48–54%) than for people over age 65 (59–65%). This is not surprising because we would expect Medicare to identify a higher proportion of individuals 65 and older given that Medicare coverage is an entitlement program that ultimately provides coverage to more than 90% of individuals aged 65 and over.

Despite the strengths of the National ALS Registry case-finding methods, it is important to recognize that

any chronic disease surveillance system in the USA will not be able to capture all cases, even for those conditions that are notifiable (e.g., cancer). Moreover, the completeness of case ascertainment for a large country like the USA (318 million in 2014) using retrospective national administrative data would be expected to be less optimal than estimates obtained from prospective studies conducted in smaller countries or in US counties or states where active prospective case-finding is feasible. Therefore, not unexpectedly, the degree of case completeness of the National ALS Registry (55–56%) was lower than that of estimates of completeness obtained with ALS registries from other countries with national healthcare systems where the data are all standardized and able to be captured within the same healthcare administration database where estimates of completeness often exceed 90 or 95% [9–14]. Nevertheless, even registries in European countries where national healthcare systems enable capture of a very high percentage of patients are pursuing the use of self-report portals and private insurance claims to improve their ascertainment of people with ALS [24].

Our findings suggest that administrative databases are an excellent foundation for identifying ALS cases for the National ALS Registry, but that additional methods should be sought to improve the completeness of case ascertainment. Although these databases do not cover the entire population of the USA, they represent a significant portion of the population most likely affected by ALS because Medicare covers individuals 65 years of age or older, and disability benefits are available to all patients with ALS even before age 65. Both the Social Security Administration (SSA) and the VA have made ALS an automatic qualification for “presumptive disability” (SSA in 2003; VA in 2008). Once a patient with ALS obtains Social Security Disability Income (SSDI), the 2-year waiting period for Medicare eligibility is waived. Although this improves the likelihood of capturing ALS in Medicare data, many younger patients are covered on private insurance (e.g., through a spouse) and may remain on that insurance coverage until death.

Medicare was the single largest contributor to case-finding, identifying 75–79% of all identified cases and uniquely identifying 49–57% of cases that were not identified by any other case-finding source. Medicare is the primary source of case ascertainment in the elderly because it is available to all US citizens aged 65 and over; however, the estimated proportion of cases missing in this age group is still 35–41%. One explanation is that a proportion of these missing cases may be categorized as “Possible ALS” and lack additional encounter data to



convert to a “Definite ALS” case. Another potential significant contributor to underascertainment was that the registry has not been allowed access to information on Medicare-eligible persons who elect to receive their care through health maintenance organizations or preferred provider organizations (i.e., “Medicare Advantage” plans). This type of Medicare coverage has been steadily increasing in recent years and now comprises approximately one-third of all Medicare-eligible persons [25]. Medicare Advantage data are now available to the registry investigators, who will use this additional case-finding method to estimate the prevalence of ALS for 2017. This will enable them to evaluate the incremental improvement in case ascertainment completeness that is achieved with Medicare Advantage and to estimate what proportion of previously underascertained cases could be attributed to not having access to Medicare Advantage data for previous annual ALS prevalence estimates.

In 2014, the size of the VA population was 19 million patients compared to 54 million in the Medicare population; therefore, we would not expect VA to contribute a large proportion of cases even though many men in the older age groups affected by ALS may have served in the military at some point in their lives. The VA source identified 19–23% of the cases in the Registry but uniquely identified only 6% of cases, indicating that most of the cases identified using this method were also ascertained by one or both of the other sources (Medicare and/or patient registry portal).

We believe that the registry methods are likely to underascertain individuals who are still employed and are covered either by private insurance or by employer-provided insurance because among individuals less than 65 years of age, there are likely to be many who are still employed (or whose spouse is employed) and getting insurance through their employer. Such individuals cannot be identified through the three case-finding sources until they either turn 65, stop working, and/or apply for Social Security Disability Income. One might propose the solution to ascertain ALS cases through private health insurance claims databases; however, employer-based health insurance claims databases cannot be obtained with any identifying information (e.g., name, age, sex, and social security number) that would enable the data to be matched to records from other case-finding sources such as Medicare or VA. As a result, it would be difficult to identify and remove duplicate ascertained patients if employer-based insurance claims data were used as an ascertainment method. Nonetheless, analysis of private insurance claims data is worthwhile. Comparing the demographic charac-

teristics of ALS cases identified through private insurance to the demographic distribution of the under 65 age group ascertained by the registry could indicate relative underrepresentation of certain demographic categories or in certain regions.

The Registry has taken several steps to add methods that will enhance volunteer web registry portal registration. These steps are expected to improve the completeness of ascertainment, especially in younger age groups as well as other groups such as those who are still employed and have not applied for disability through Medicare. These methods include partnering with ALS advocacy groups and patient service organizations (ALSA, MDA) to increase Registry awareness among patients and caregivers. As the primary provider of medical care to low-income individuals, Medicaid would appear to be a desirable method of case ascertainment; however, in our earlier study during 2002–2004 of ALS prevalence in the USA [19], only 3% of cases were uniquely identified by Medicaid, suggesting that not using Medicaid data might have limited impact on the completeness of case ascertainment. In the past registry years, Medicaid could not feasibly be used as a timely source of data because it typically has taken 3–4 years for CMS to receive Medicaid data from all states. Efforts are now being made by CMS to make national Medicaid data available on a more rapid timetable, and Medicaid can be reconsidered in the future as a case-finding method if timeliness improves. Another approach would be to conduct a study to determine how many patients with ALS are seen in safety net clinics such as Federally Qualified Health Clinics (FQHCs); this might allow access to personal identifying information for deduplication purposes since those clinics are run by the HRSA (Health Resources and Services Organization), a government agency. However, the majority of patients seen at FQHCs have Medicaid coverage, and therefore only one of the two methods would need to be used to enhance the identification of low-income individuals.

## Discussion/Conclusion

Information provided by our analysis suggests that the methods used by the National ALS Registry have provided an excellent methodological foundation for estimating national ALS prevalence in a large country with a fragmented healthcare system and may be valuable for tracking cases of other neurological diseases and estimating their prevalence in the USA. Additional methods are needed to improve the completeness of case ascertain-

ment and to achieve better representation of younger patients, those who are still employed, and those who receive medical care in safety net clinics.

## Acknowledgments

We gratefully acknowledge the people who assisted with the collection of data for the National ALS Registry and who structured the analytic data set prior to data analysis including Ms. Marchelle Sanchez and Dr. Oleg Muravov.

## Statement of Ethics

All activities involving the National ALS Registry have been reviewed and approved by the Institutional Review Board (IRB) of the Centers for Disease Control and Prevention (CDC)/ATSDR.

## Conflict of Interest Statement

Dr. Mehta, Dr. Horton, and Jaime Raymond are employees of the federal government. Dr. Wendy Kaye is employed by the McKing Consulting Corporation and is a paid contractor for the National ALS Registry. There are no patents, products in development, or marketed products to declare. There are no other competing interests for the submitted research.

## Funding Sources

This project was designed, implemented, and funded by the federal government of the USA. The funder provided support in the form of salaries for P.M., J.R., and D.K.H. and in the form of

salary support for L.M.N., B.T., and T.W. via a contract with the CDC. The McKing Consulting Corporation provided support in the form of salary for W.K. but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Author Contributions

Lorene M. Nelson, PhD, MS: Design and conceptualization of the study and drafted the manuscript. Barbara Topol, MS: Acquired data, analyzed data, and revised the manuscript for intellectual content. Wendy Kaye, PhD: Design and conceptualization of the study, acquired data, and revised the manuscript for intellectual content. Jaime Raymond, MPH: Acquired data, analyzed data, and revised the manuscript for intellectual content. D. Kevin Horton, DrPH, MSPH: Acquired data and revised the manuscript for intellectual content. Paul Mehta, MD: Design and conceptualization of the study, acquired data, and revised the manuscript for intellectual content. Todd Wagner, PhD: Design and conceptualization of the study and revised the manuscript for intellectual content.

## Data Availability Statement

The data that support the findings of this study are not publicly available due to the privacy risks of research subjects and to the policies of the data providers (Medicare, Medicaid, and Veterans Administration) and the Office of Management and Budget (OMB). The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the Agency for Toxic Substances and Disease Registry.

## References

- 1 Chiò A, Mora G, Calvo A, Mazzini L, Bottacchi E, Mutani R, et al. Epidemiology of ALS in Italy: a 10-year prospective population-based study. *Neurology*. 2009 Feb 24;72(8):725–31.
- 2 Donaghy C, Clarke J, Patterson C, Kee F, Hardiman O, Patterson V. The epidemiology of motor neuron disease in Northern Ireland using capture-recapture methodology. *Amyotroph Lateral Scler*. 2010;11:374–8.
- 3 Huisman MH, de Jong SW, van Doormaal PT, Weinreich SS, Schelhaas HJ, van der Kooij AJ, et al. Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. *J Neurol Neurosurg Psychiatry*. 2011 Oct;82(10):1165–70.
- 4 Golby R, Poirier B, Fabros M, Cragg JJ, Yousefi M, Cashman N. Five-year incidence of amyotrophic lateral sclerosis in British Columbia (2010–2015). *Can J Neurol Sci*. 2016 Nov;43(6):791–5.
- 5 Rosenbohm A, Peter RS, Erhardt S, Lulé D, Rothenbacher D, Ludolph AC, et al. Epidemiology of amyotrophic lateral sclerosis in Southern Germany. *J Neurol*. 2017 Apr;264(4):749–57.
- 6 US Public Health Service. ALS registry act. Washington, DC: 110th Congress. Public Law 2008;122 Stat 4047:110–373.
- 7 21st Century Cures Act, enacted by the 114th United States Congress in December 2016. An act to accelerate the discovery, development, and delivery of 21st century cures, and for other purposes. *PUBLIC LAW* 114–255: DEC. 13, 2016 130 STAT. 1033 Public Law 114–255 114th Congress An Act. Available from: <https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf>.
- 8 Kaye WE, Sanchez M, Wu J. Feasibility of creating a National ALS Registry using administrative data in the United States. *Amyotroph Lateral Scler Frontotemporal Degener*. 2014 Sep;15(5–6):433–9.
- 9 Turabelidze G, Zhu BP, Schootman M, Malone JL, Horowitz S, Weidinger J, et al. An epidemiologic investigation of amyotrophic lateral sclerosis in Jefferson County, Missouri, 1998–2002. *Neurotoxicology*. 2008 Jan;29(1):81–6.
- 10 Jordan H, Fagliano J, Rechtman L, Lefkowitz D, Kaye W. Population-based surveillance of amyotrophic lateral sclerosis in New Jersey, 2009–2011. *Neuroepidemiology*. 2014;43(1):49–56.
- 11 Wagner L, Rechtman L, Jordan H, Ritsick M, Sanchez M, Sorenson E, et al. State and metropolitan area-based amyotrophic lateral sclerosis (ALS) surveillance. *Amyotroph Lateral Scler Frontotemporal Degener*. 2015;17(1–2):128–34.

- 12 Freer C, Hylton T, Jordan HM, Kaye WE, Singh S, Huang Y. Results of Florida's amyotrophic lateral sclerosis surveillance project, 2009–2011. *BMJ Open*. 2015 Apr 21;5(4):e007359.
- 13 Valle J, Roberts E, Paulukonis S, Collins N, English P, Kaye W. Epidemiology and surveillance of amyotrophic lateral sclerosis in two large metropolitan areas in California. *Amyotroph Lateral Scler Frontotemporal Degener*. 2015 Jun;16(3–4):209–15.
- 14 Jordan H, Rechtman L, Wagner L, Kaye WE. Amyotrophic lateral sclerosis surveillance in Baltimore and Philadelphia. *Muscle Nerve*. 2015 Jun;51(6):815–21.
- 15 Kaye WE, Wagner L, Wu R, Mehta P. Evaluating the completeness of the national ALS registry, United States. *Amyotroph Lateral Scler Frontotemporal Degener*. 2018 Feb;19(1–2):112–7.
- 16 Wittes JT, Colton T, Sidel VW. Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *J Chronic Dis*. 1974 Feb;27(1):25–36.
- 17 Hook EB, Regal RR. The value of capture-recapture methods even for apparent exhaustive surveys. The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. *Am J Epidemiol*. 1992;135:1060–7.
- 18 Civilian Health and Medical Program of the Department of Veterans Affairs (CHAMPVA). Available from: <https://www.va.gov/communitycare/programs/dependents/champva/index.asp> Accessed 2021 Jul 28.
- 19 Nelson LM, Topol B, Kaye W, Williamson D, Horton DK, Mehta P, et al. Estimation of the prevalence of amyotrophic lateral sclerosis in the United States using national administrative healthcare data from 2002 to 2004 and capture-recapture methodology. *Neuroepidemiology*. 2018;51(3–4):149–57.
- 20 SAS Institute Inc., Cary, NC.
- 21 Chao A. *Capture-recapture for human populations*. Wiley StatsRef: statistics reference online. New York: John Wiley & Sons, Ltd; 2014–2015.
- 22 Bishop YMM, Fienberg SE, Holland PW. *Discrete multivariate analysis: theory and practice*. Cambridge: MIT Press; 1975. p. 229–56.
- 23 Gimenez O, Choquet R, Lamor L, Scofield P, Fletcher D, LeBreton J-D, et al. Efficient profile-likelihood confidence intervals for capture-recapture models. *J Agricult Biolog and Env Statistics*. 2005;10(2):184–96.
- 24 Opie-Martin S, Ossher L, Bredin A, Kulka A, Pearce N, Talbot K, et al. Motor neuron disease register for England, Wales and Northern Ireland-an analysis of incidence in England. *Amyotroph Lateral Scler Frontotemporal Degener*. 2021;22(1–2):86–93.
- 25 Kaiser Family Foundation. An overview of medicare. Available from: <https://www.kff.org/medicare/issue-brief/an-overview-of-medicare/> Accessed 2021 Mar 4.