Estimation of prevalence of autoimmune diseases in the United States using electronic health record data

Aaron H. Abend,[1] Ingrid He,[1] Neil Bahroos,[2] Stratos Christianakis,[2] Ashley B. Crew,[3] Leanna M. Wise,[2] Gloria P. Lipori,[4] Xing He,[4] Shawn N. Murphy,[5] Christopher D. Herrick,[6] Jagannadha Avasarala,[7] Mark G. Weiner,[8] Jacob S. Zelko,[9] Erica Matute-Arcos,[1] Mark Abajian,[2] Philip R.O. Payne,[10] Albert M. Lai,[10] Heath A. Davis,[11] Asher A. Hoberg,[11] Chris E. Ortman,[11] Amit D. Gode,[12] Bradley W. Taylor,[12] Kristen I. Osinski,[12] Damian N. Di Florio,[13,14,15] Noel R. Rose,[16] Frederick W. Miller,[17] George C. Tsokos,[18] and DeLisa Fairweather[13,14,19]

[1]Autoimmune Registry Inc., Guilford, Connecticut, USA. [2]Division of Bioinformatics, Department of Population and Public Health Sciences, Keck School of Medicine and [3]Department of Dermatology, Keck School of Medicine, University of Southern California, Los Angeles, California, USA. [4]University of Florida Health, Gainesville, Florida, USA. [5]Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, USA. [6]Research Information Systems and Computing, Mass General Brigham, Somerville, Massachusetts, USA. [7]Department of Neurology, University of Kentucky Medical Center, Lexington, Kentucky, USA. [8]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA. [9]Georgia Tech Research Institute, Atlanta, Georgia, USA. [10]Institute for Informatics, Data Science, and Biostatistics, Washington University School of Medicine, St. Louis, Missouri, USA. [11]Institute for Clinical and Translational Science, University of Iowa, Iowa City, Iowa, USA. [12]Clinical & Translational Science Institute, Medical College of Wisconsin, Milwaukee, Wisconsin, USA. [13]Department of Cardiovascular Medicine, Mayo Clinic, Jacksonville, Florida, USA. [14]Center for Clinical and Translational Science and [15]Mayo Clinic Graduate School of Biomedical Sciences, Mayo Clinic, Rochester, Minnesota, USA. [16]Harvard Medical School, Boston, Massachusetts, USA. [17]National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina, USA. [18]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA. [19]Department of Immunology, Mayo Clinic, Jacksonville, Florida, USA.

BACKGROUND. Previous epidemiologic studies of autoimmune diseases in the US have included a limited number of diseases or used metaanalyses that rely on different data collection methods and analyses for each disease.

METHODS. To estimate the prevalence of autoimmune diseases in the US, we used electronic health record data from 6 large medical systems in the US. We developed a software program using common methodology to compute the estimated prevalence of autoimmune diseases alone and in aggregate that can be readily used by other investigators to replicate or modify the analysis over time.

RESULTS. Our findings indicate that over 15 million people, or 4.6% of the US population, have been diagnosed with at least 1 autoimmune disease from January 1, 2011, to June 1, 2022, and 34% of those are diagnosed with more than 1 autoimmune disease. As expected, females (63% of those with autoimmune disease) were almost twice as likely as males to be diagnosed with an autoimmune disease. We identified the top 20 autoimmune diseases based on prevalence and according to sex and age.

CONCLUSION. Here, we provide, for what we believe to be the first time, a large-scale prevalence estimate of autoimmune disease in the US by sex and age.

FUNDING. Autoimmune Registry Inc., the National Heart Lung and Blood Institute, the National Center for Advancing Translational Sciences, the Intramural Research Program of the National Institute of Environmental Health Sciences.

Introduction

Autoimmune diseases are a diverse group of chronic inflammatory pathologies marked by a dysfunctional innate and adaptive immune system after exposure to proinflammatory environmental agents resulting in subsequent end-organ damage that lead to clinical disease manifestations (1). Although studies of the prevalence and incidence of individual autoimmune diseases have been reported, the prevalence of autoimmune diseases as a class has only been estimated 5 times to date, most recently in 2023 in the United Kingdom (Table 1) (2–6). Many challenges exist to obtain accurate data on the prevalence of all autoimmune diseases, including the lack of an international consensus on the definition of autoimmune disease and which specific entities fall into this category (7).

Precedence for classifying diseases into major categories can been seen in cancer (8), cardiovascular diseases (9), and organ-specific diseases including the skin (10), respiratory (11), and digestive systems (12). Prevalence statistics for individual diseases provide context to interpret test results used to diagnose patients (13). Prevalence

▶ Related Editor's Note: https://doi.org/10.1172/JCI190119

Authorship note: AHA and DF contributed equally to this work.
Conflict of interest: The authors have declared that no conflict of interest exists.

Submitted: February 1, 2024; **Accepted:** December 10, 2024; **Published:** December 12, 2024.
Reference information: *J Clin Invest*. 2025;135(4):e178722.
https://doi.org/10.1172/JCI178722.

Table 1. Prior studies of autoimmune disease prevalence

Author (Reference)	Date	No. of Autoimmune Diseases	Approach	Location	Published prevalence	Extension to 2022 US Population (ref. 35)
Jacobson (ref. 2)	1996	24	Meta-analysis	Worldwide	3.2%	10,620,243
Eaton (ref. 3)	2006	30	EHR data analysis	Denmark	4%	13,300,000
Cooper (ref. 4)	2009	29	Update of Jacobson using Eaton research	Worldwide	7.6%–9.4%	25,000,000–31,000,000
Hayter (ref. 5)	2012	81	Meta-analysis	Worldwide	4.5%	14,962,500
Conrad (ref. 6)	2023	19	EHR data analysis	United Kingdom	10.2%	33,966,000

statistics by disease class can help assess the burden of these diseases on a population. There is a need to assess the prevalence of autoimmune diseases as a class to fully appreciate their impact on society, where many rare autoimmune conditions may otherwise be ignored.

Knowledge of disease class prevalence (i.e., autoimmune diseases) is also important to raise public awareness of autoimmune diseases in general, which helps channel funding to individual autoimmune diseases and assists in the recognition of rare autoimmune diseases. As highlighted in a recent National Academy of Sciences, Engineering, and Medicine report, research and public awareness efforts for autoimmune diseases have focused almost exclusively on a limited number of autoimmune diseases including inflammatory bowel disease, multiple sclerosis, type I diabetes, rheumatoid arthritis (RA), and systemic lupus erythematosus (14). The American Heart Association (AHA) describes deaths and other outcomes for all cardiovascular diseases before breaking data down by categories of heart disease (15), and these data are used by the AHA for public awareness campaigns to emphasize the importance of clinical and research efforts to decrease heart disease. Similarly, prevalence and incidence data on cancer as a class provided by the National Cancer Institute's Surveillance, Epidemiology and End Results Program (SEER, https://seer.cancer.gov) and the American Cancer Society (https://www.cancer.gov) are used to emphasize the need to research cures for cancer. Thus, knowing the overall prevalence of diseases by class is an important component of research and public health awareness efforts in the US. The research community and the public should have access to similar data for autoimmune diseases in the US as they do for heart disease and cancer. This study is the first to our knowledge to examine a large number of autoimmune diseases in the US using nationwide data.

Another reason to gather information on autoimmune diseases as a group is that, due to shared environmental or genetic risk factors, individuals quite frequently suffer from multiple autoimmune conditions (16, 17). For example, polymorphisms in certain immune genes have been found to occur in several autoimmune diseases (18), which provides a possible explanation for the occurrence of multiple autoimmune diseases in the same individual. Research

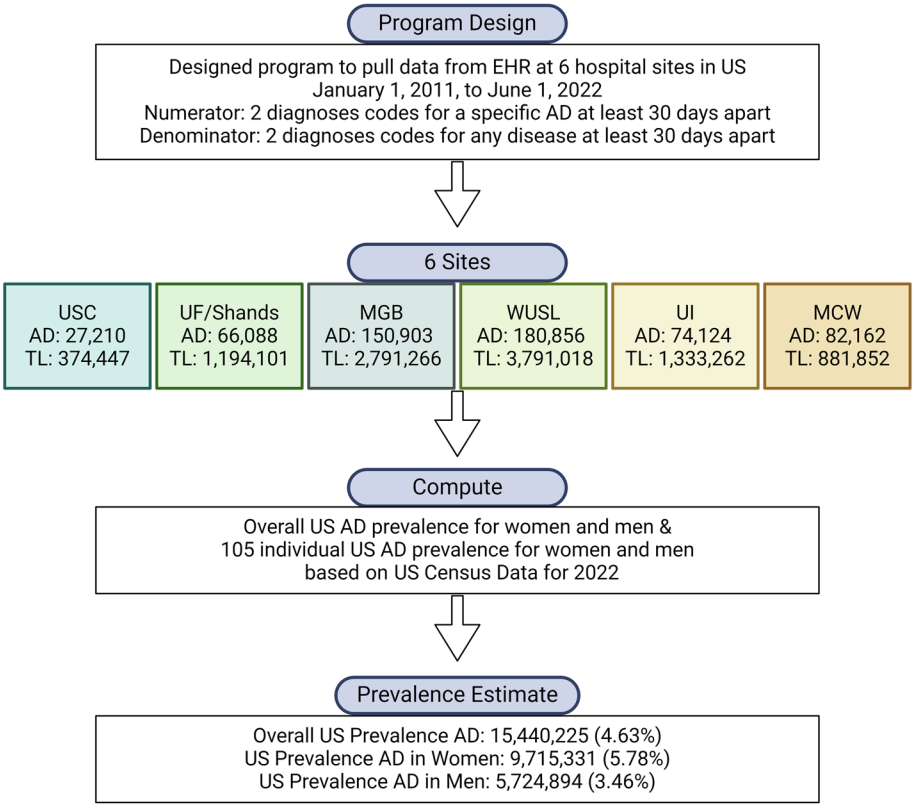

Figure 1. A flow chart of the study design. A total (TL) of 10,365,946 individuals were identified from the electronic health record (EHR) from January 1, 2011, to June 1, 2022, from 6 healthcare sites in the US based on a program that identified patients with 2 diagnoses codes for any disease at least 30 days apart (denominator). From this total, 581,343 individuals were identified with 1 of 105 specific autoimmune diseases (ADs) based on 2 diagnoses codes at least 30 days apart (numerator) in the EHR. Overall AD prevalence for women and men was computed based on US Census Data for 2022. The 6 healthcare sites included University of Southern California (USC), University of Florida (UF)/ Shands, Mass General Brigham (MGB), Washington University of St. Louis (WUSL), University of Iowa (UI), and the Medical College of Wisconsin (MCW). The image was designed using BioRender.

Table 2. Estimated prevalence of autoimmune disease by sex and age

		Autoimmune Disease Counts /Denominators (%)			US Autoimmune Disease Prevalence[A]		
	Age	Female	Male	Total	Female	Male	Total
University of Southern California (USC)	0–17	18/3,571 (0.50%)	23/4,098 (0.56%)	41/7,669 (0.53%)	177,792	207,964	385,755
	18–44	4,542/53,306 (8.52%)	2,350/43,826 (5.36%)	6,892/97,132 (7.10%)	5,047,816	3,286,620	8,334,435
	45–64	6,351/58,371 (10.88%)	3,025/52,288 (5.79%)	9,376/110,659 (8.47%)	4,531,403	2,369,445	6,900,847
	≥65	7,189/78,373 (9.17%)	3,712/80,614 (4.60%)	10,901/158,987 (6.86%)	2,925,910	1,193,742	4,119,652
	Total	**18,100 /193,621 (9.35%)**	**9,110 /180,826 (5.04%)**	**27,210 /374,447 (7.27%)**	**12,682,920**	**7,057,769**	**19,740,690**
University of Florida and Shands Health System (UF/Shands)	0–17	1,938/97,252 (1.99%)	1,496/111,593 (1.34%)	3,434/208,845 (1.64%)	702,885	496,736	1,199,622
	18–44	14,580/255,343 (5.71%)	6,330/162,792 (3.89%)	20,910/418,135 (5.00%)	3,382,720	2,383,329	5,766,049
	45–64	15,570/162,722 (9.57%)	6,027/127,111 (4.74%)	21,597/289,833 (7.45%)	3,985,016	1,941,964	5,926,980
	≥65	13,633/144,278 (9.45%)	6,514/133,010 (4.90%)	20,147/277,288 (7.27%)	3,014,049	1,269,627	4,283,676
	Total	**45,721/659,595 (6.93%)**	**20,367/534,506 (3.81%)**	**66,088/1,194,101 (5.53%)**	**11,084,669**	**6,091,657**	**17,176,327**
Mass General Brigham (MGB)	0–17	1,656/129,636 (1.28%)	1,328/144,779 (0.92%)	2,984/274,415 (1.09%)	450,572	339,879	790,450
	18–44	26,077/530,203 (4.92%)	12,380/357,965 (3.46%)	38,457/888,168 (4.33%)	2,913,720	2,119,793	5,033,514
	45–64	33,517/442,193 (7.58%)	15,663/331,135 (4.74%)	49,180/773,328 (6.36%)	3,156,756	1,937,283	5,094,039
	≥65	39,626/475,013 (8.34%)	20,656/380,342 (5.43%)	60,282/855,355 (7.05%)	2,660,932	1,407,942	4,068,874
	Total	**100,876/1,577,045 (6.40%)**	**50,027/1,214,221 (4.12%)**	**150,903/2,791,266 (5.41%)**	9,181,979	5,804,897	14,986,876
Washington University of St. Louis (WUSL)	0–17	1985/205010 (0.97%)	1625/230543 (0.70%)	3,610/435,553 (0.83%)	341,519	261,176	602,694
	18–44	22356/530,203 (3.62%)	12298/523314 (2.35%)	34,654/1,141,502 (3.04%)	2,142,427	1,440,408	3,582,835
	45–64	35918/493236 (7.28%)	15324/382087 (4.01%)	51,242 /875,323 (5.85%)	3,032,809	1,642,605	4,675,414
	≥65	61299/728169 (8.42%)	30051/30051 (4.92%)	91,350 /1,338,640 (6.82%)	2,685,222	1,276,165	3,961,386
	Total	**121,558 /2,044,603 (5.95%)**	**59,298 /1,746,415 (3.40%)**	**180,856 /3,791,018 (4.77%)**	**8,201,976**	**4,620,353**	**12,822,330**
University of Iowa (UI)	0–17	1,192/50,883 (2.34%)	987/62,531 (1.58%)	2,179/113,414 (1.92%)	826,290	584,862	1,411,152
	18–44	10,010/218,309 (4.59%)	6,056/185,676 (3.26%)	16,066/403,985 (3.98%)	2,716,407	1,999,141	4,715,549
	45–64	11,832/180,616 (6.55%)	6,519/145,202 (4.49%)	18,351/325,818 (5.63%)	2,728,284	1,838,788	4,567,072
	≥65	23,979/257,972 (9.30%)	13,549/232,073 (5.84%)	37,528/490,045 (7.66%)	2,964,951	1,513,545	4,478,496
	Total	**47,013/707,780 (6.64%)**	**27,111/625,482 (4.33%)**	**74,124/1,333,262 (5.56%)**	**9,235,933**	**5,936,336**	**15,172,269**
Medical College of Wisconsin (MCW)	0–17	186/23,373 (0.80%)	138/25,029 (0.55%)	324/48,402 (0.67%)	280,690	204,299	484,990
	18–44	11,202 /158,450 (7.07%)	5,271 /112,724 (4.68%)	16,473 /271,174 (6.07%)	4,188,281	2,866,091	7,054,372
	45–64	17,541 /137,795 (12.73%)	7,938 /110,218 (7.20%)	25,479 /248,013 (10.27%)	5,301,620	2,949,727	8,251,346
	≥65	26,301 /172,517 (15.25%)	13,585 /141,746 (9.58%)	39,886 /314,263 (12.69%)	4,862,945	2,484,629	7,347,574
	Total	**55,230 /492,135 (11.22%)**	**26,932 /389,717 (6.91%)**	**82,162 /881,852 (9.32%)**	**14,633,536**	**8,504,747**	**23,138,282**
Combined Sites	0–17	6,975/509,725 (1.37%)	5,597/578,573 (0.97%)	12,572/1,088,298 (1.16%)	482,656	358,450	841,106
	18–44	88,767/1,833,799 (4.84%)	44,685/1,386,297 (3.22%)	133,452/3,220,096 (4.14%)	2,867,690	1,975,690	4,843,381
	45–64	120,729/1,474,933 (8.19%)	54,496/1,148,041 (4.75%)	175,225/2,622,974 (6.68%)	3,409,000	1,944,153	5,353,153
	≥65	172,027/1,856,322 (9.27%)	88,067/1,578,256 (5.58%)	260,094/3,434,578 (7.57%)	2,955,985	1,446,600	4,402,586
	Total	**388,498/5,674,779 (6.85%)**	**192,845/4,691,167 (4.11%)**	**581,343/10,365,946 (5.61%)**	**9,715,331 (5.78%)**	**5,724,894 (3.46%)**	**15,440,225 (4.63%)**

[A]Based-on US Census Data for 2022 (found in Supplemental Table 3).

strategies that count individual autoimmune diseases and then aggregate those statistics for multiple autoimmune diseases count individuals more than once and thereby might overstate prevalence. This is a common issue found in metadata assessments of prevalence that makes estimations of prevalence over time difficult.

Finally, there is recent evidence suggesting that the prevalence of biomarkers, such as antinuclear antibodies (19), has increased for at least some autoimmune diseases, and the scientific community and the public need to know whether this increase is associated with a parallel increase in the incidence and prevalence of autoimmune diseases (20, 21). There is an urgency to develop approaches to compute the prevalence of autoimmune diseases that can be replicated longitudinally. Therefore, we aimed to provide an update on the prior estimates by providing a current overall autoimmune disease prevalence estimate in the US according to sex and age as well as for 105 autoimmune diseases using electronic health record (EHR) data.

Results

In order to determine the prevalence of autoimmune diseases in the US we selected 105 diseases that were listed in the textbook "The Autoimmune Diseases" by Rose and MacKay, 5th Edition (1), that had substantiative evidence of an autoimmune pathology (Supplemental Table

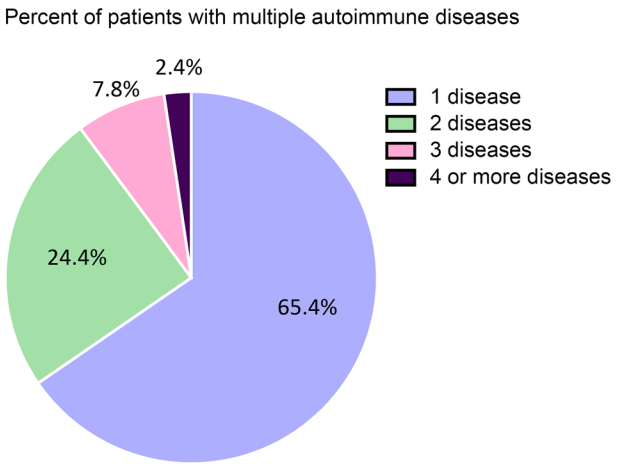

Figure 2. Prevalence of multiple autoimmune diseases. Individuals with 1 autoimmune disease are known to often suffer from another autoimmune condition. Research strategies that count individual autoimmune diseases and then aggregate those statistics for multiple autoimmune diseases count individuals more than once and thereby might overstate prevalence. This figure reports the frequency of multiple autoimmune diseases in this study, indicating that this could be an issue in certain prevalence estimates and indicates how often they cooccur.

1; supplemental material available online with this article; https://doi.org/10.1172/JCI178722DS1). Our list included all autoimmune diseases for which there is evidence in the literature that self-reactive T cells and/or antibodies contribute to or cause the disease (22). We included cis-females (referred to hereafter as females) and cis-males (referred to hereafter as males) with no age restriction, according to gender information provided in the EHR. A list of diseases that lack clear evidence for an autoimmune pathology, but are often categorized or described as autoimmune, is included in Supplemental Table 2, along with published and computed prevalence for thoroughness; however, these conditions were not included in our prevalence estimates.

Between January 1, 2011 and June 1, 2022, we identified a total of 581,343 individuals from 6 medical systems across the US serving a population of 10,365,946 that were diagnosed with at least 1 of the 105 autoimmune diseases considered in this study (Figure 1 and Table 2). Extending these statistics to an estimated US population of 333.3 million in 2022 (23) (Supplemental Table 3) gives an overall computed prevalence of 15,440,225 individuals (95% CI 15,437,949–15,442,501), or 4.6% of the US population with an autoimmune disease. The prevalences of each of the 105 individual autoimmune diseases by sex are shown in Supplemental Table 1, along with the published estimated prevalence as of January 1, 2022. For 22 of the 105 diseases, there were no patients who met the requirements for inclusion, and there were 9 diseases for which the patient counts were below 10 and therefore estimates have not been reported. The overall estimated prevalence for females was 9,715,331 (95% CI 9,680,412–9,750,250) or 5.8% of the US female population, and for males was 5,724,894 (95% CI 5,695,208–5,754,578) or 3.5% of the US male population (Table 2).

Table 3. Top 20 most prevalent autoimmune diseases

Rank	Autoimmune Disease	Computed Estimated US Prevalence			Female Ratio	Rate/100,000
		Female	Male	Total[A]		
1	Rheumatoid arthritis	1,827,271	653,179	2,480,449	74%	744.2
2	Psoriasis	1,065,966	1,005,908	2,071,875	51%	621.6
3	Diabetes mellitus type 1	894,091	982,002	1,876,093	48%	562.9
4	Graves' disease	1,293,040	415,444	1,708,484	76%	512.6
5	Autoimmune thyroiditis	1,058,454	187,061	1,245,515	85%	373.7
6	Crohn's disease	622,853	574,725	1,197,578	52%	359.3
7	Multiple sclerosis	809,019	325,368	1,134,387	71%	340.4
8	Systemic lupus erythematosus	860,667	131,187	991,854	87%	297.6
9	Ulcerative colitis	464,741	483,672	948,413	49%	284.6
10	Sjögren's disease	545,176	79,187	624,363	87%	187.3
11	Celiac disease	393,901	173,765	567,666	69%	170.3
12	Polymyalgia rheumatica	304,398	202,417	506,815	60%	152.1
13	Autoimmune gastritis	171,229	100,441	271,670	63%	81.5
14	Vitiligo	130,263	120,299	250,562	52%	75.2
15	Autoimmune thrombocytopenic purpura	130,821	117,297	248,118	53%	74.4
16	Aplastic anemia	116,647	122,113	238,761	49%	71.6
17	Alopecia areata	130,762	97,928	228,690	57%	68.6
18	Juvenile rheumatoid arthritis	155,002	69,415	224,417	69%	67.3
19	Systemic sclerosis	180,092	40,344	220,435	82%	66.1
20	Autoimmune hepatitis	146,491	45,090	191,582	76%	57.5

[A]Prevalence order is based on total US prevalence column.

Table 4. Autoimmune diseases with the highest percentage in females

Rank	Disease	Computed Estimated US Prevalence			Female Ratio[A]	Rate/100,000
		Female	Male	Total		
1	Lichen sclerosis	119,112	6,596	125,708	96%	86.8
2	Sjögren's disease	545,176	79,187	624,363	89%	293.2
3	Systemic lupus erythematosus	860,667	131,187	991,854	89%	27.4
4	Primary biliary cholangitis	79,408	13,401	92,810	88%	368.4
5	Autoimmune thyroiditis	1,058,454	187,061	1,245,515	87%	4.2
6	Systemic sclerosis	180,092	40,344	220,435	84%	65.2
7	Cutaneous lupus erythematosus	69,255	15,705	84,959	84%	26.9
8	SLE glomerulonephritis syndrome	73,833	16,926	90,759	84%	28.2
9	Autoimmune hepatitis	146,491	45,090	191,582	79%	56.7
10	Graves' disease	1,293,040	415,444	1,708,484	79%	506.1
11	Dermatomyositis	81,521	27,466	108,987	78%	30.4
12	Rheumatoid vasculitis	5,576	1,919	7,495	78%	9.0
13	Neuromyelitis optica	22,420	7,887	30,307	77%	735.0
14	Rheumatoid arthritis	1,827,271	653,179	2,480,449	77%	336.3
15	Vogt-Koyanagi-Harada disease	3,404	1,221	4,625	77%	11.1
16	Microscopic polyangiitis	3,463	1,326	4,789	76%	44.9
17	Multiple sclerosis	809,019	325,368	1,134,387	75%	168.3
18	Behçet's syndrome	26,587	10,993	37,580	74%	66.6
19	Antiphospholipid syndrome	106,699	44,706	151,405	74%	3.1
20	Temporal arteritis	118,056	52,000	170,056	73%	1.5

[A]Order based on female sex ratio. SLE, systemic lupus erythematosus.

As expected, most patients diagnosed with autoimmune diseases were female (63%) compared to male (37%) for an overall sex ratio of 1.7:1 female-to-male. The number of individuals not reporting a sex of male or female was under 20 for many diseases at many of the sites in this study and so is not reported to protect patient privacy. Additionally, 65% of patients had 1 autoimmune disease whereas 24% had 2, 8% had 3, and 2% had 4 or more autoimmune diseases (Figure 2 and Supplemental Table 4). The top 20 autoimmune diseases based on prevalence are listed in Table 3 with RA, psoriasis, type I diabetes mellitus, Graves' disease, and autoimmune thyroiditis being the top 5. Interestingly, 17 of the top 20 autoimmune diseases occurred more often in females than males. The top autoimmune diseases in females or males based on sex ratio are listed in Tables 4 and 5, respectively.

Discussion

In this study we developed a new tool to estimate the prevalence of autoimmune disease in the US. Our methodology offers the following advantages: (a) The entire analysis can be run at any site that has data in the widely used Observational Medical Outcomes Partnership (OMOP) model; (b) the tool is easy to use and generally takes a few hours to run; (c) the diseases selected for inclusion can be easily modified; and (d) the tool can be modified to add additional parameters such as medications and labs, to improve diagnostic specificity and sensitivity and for individual research purposes.

Our estimate of over 15 million, or 4.6%, individuals with autoimmune disease in the US is below some commonly quoted estimates of 7%–10% (see Table 1). Our selection criteria that required 2 diagnosis codes at least 30 days apart aimed to reduce counting individuals that were being investigated but had not yet been diagnosed with an autoimmune disease. Several estimates of the prevalence of autoimmune disease have been aggregates of meta-analyses of disease-specific prevalence data, which likely over estimates prevalence due to double counting, since a patient with more than 1 disease is counted in more than 1 of the disease estimates.

We confirmed an overall sex ratio for autoimmune diseases of around 1.7:1 female-to-male. We have provided prevalence by sex for 105 individual autoimmune diseases (Supplemental Table 1). These data are needed to better understand the impact of individual autoimmune diseases by biological sex and to support the need for clinical and basic research examining overall autoimmune and disease-specific mechanisms. Research in autoimmune diseases has not kept pace with advances in other disease categories like cancer and heart disease because of relatively lower funding levels and a paucity of specific data for the US population.

Several previous studies found that patients with 1 autoimmune disease are more likely to develop another autoimmune disease (16, 17). However, there has been a lack of data on the prevalence of cooccurrence of autoimmune diseases overall in the US. In this study we show that as many as 24% of patients are diagnosed with 2 autoimmune diseases and 2% have 4 or more autoimmune diseases concurrently. More research is needed to understand which autoimmune diseases cooccur and if common mechanisms can be targeted with improved diagnostic tests and therapies.

Table 5. Autoimmune diseases with the highest percentage in males

Rank	Disease	Computed Estimated US Prevalence			Female Ratio[A]	Rate/100,000
		Female	Male	Total		
1	Acquired hemophilia	205	872	1,078	19%	0.3
2	Inclusion body myositis	7,776	17,450	25,226	31%	7.6
3	Reactive arthritis	10,858	21,707	32,565	33%	9.8
4	Idiopathic pulmonary fibrosis	55,873	111,050	166,923	33%	50.1
5	Acute febrile mucocutaneous lymph node syndrome	18,077	34,167	52,243	35%	15.7
6	Primary sclerosing cholangitis	21,363	36,051	57,414	37%	17.2
7	Chronic inflammatory demyelinating polyradiculoneuropathy	32,603	50,674	83,277	39%	25.0
8	Autoimmune lymphoproliferative syndrome	205	314	519	40%	0.2
9	Pure red cell aplasia	3,932	5,863	9,795	40%	2.9
10	Postmyocardial infarction syndrome	1,702	2,478	4,180	41%	1.3

[A]Order based on female sex ratio.

There are a number of limitations to our study. The use of EHR data to determine who has an autoimmune disease is complicated by several factors. Since the diagnosis of a given autoimmune disease is rarely, if ever, contingent only on the presence of clear biomarkers, autoimmune disease codes in the EHR might not be accurate (24). Many patients have diagnoses that are subsequently refined or completely changed as their symptoms and clinical findings evolve (25–27). Some diseases can be caused by autoimmune or nonautoimmune processes. An example would be the diagnosis of type 1 diabetes mellitus in a patient who has undergone a total pancreatectomy (28). We could also miss patients with a single diagnosis code since we only count patients with at least 2 diagnosis codes. It is also known that autoimmune diseases evolve over time and involve nonspecific clinical signs and symptoms that can mimic other diseases that may result in an underdiagnosis of many of these diseases. Rare diseases, such as antisynthetase syndrome and IgG4-related disease, lack specific ICD-10 codes (29). Though our analysis uses Systematized Nomenclature of Medicine (SNOMED) codes, which do exist for these diseases, we know that EHR data at the sites we studied, which use ICD-10 coding, will not identify these patients. Using broader disease names, such as "myositis" to capture antisynthetase syndrome, however, captures too many patients who do not have this autoimmune disease. Therefore, we included a category of "Autoimmune Disease Not Otherwise Specified" which will capture some of these diseases. Additionally, our dataset was based on data from academic medical centers. Such systems include more specialists and fewer general practitioners, leading to possible selection bias. Coding error is another limitation: type 2 diabetics are often miscoded for type 1, leading to inflated values for that condition. Another limitation is that patient death is not typically recorded consistently in EHR systems, so patients who died during the study period will be counted in the numerator. Since these patients are also included in the denominator, this limitation should not have a significant impact on overall prevalence statistics. Also in the US, individuals move location frequently and so it is possible that the same patient could be counted at more than 1 location. However, the 6 sites in this study are in diverse locations which should reduce this error. In spite of these limitations, we believe that the use of a common data model and methodology for all conditions provides support for the accuracy of our estimate, and the software used to compute our estimates can be improved over time as these many limitations are addressed. And, finally, a number of the conditions in our list of autoimmune diseases may not be considered by all investigators to have sufficient evidence to name them autoimmune diseases. For a conservative approach, we included diseases discussed in the textbook "The Autoimmune Diseases" edited by Rose and Mackay (1). However, we fully acknowledge that some conditions may be considered 'autoinflammatory' or simply inflammatory conditions. Our goal was to provide data on prevalence by sex for individual autoimmune diseases that may help move the field forward in order to better address these and other issues in the field. Our development of a relatively simple tool now made available freely to the clinical and research community will hopefully fulfill this goal.

Conclusions. We developed a new analysis tool to determine the overall and individual prevalence of autoimmune diseases in the US or other countries. Using this tool and data from the EHR of 6 major medical systems in the US, we estimated that autoimmune disease affected over 15 million individuals in the US in 2022, which is 4.6% of the population. Females represented 63% of those with autoimmune disease, and males 37%, a sex ratio of 1.7:1 female-to-male. We report high levels of comorbid autoimmune diseases with 24% of patients with autoimmune disease diagnosed with 2 autoimmune diseases and 8% with 3. Accurate data on the prevalence of autoimmune diseases as a category of disease and for individual autoimmune diseases are needed to further clinical and basic research to improve diagnosis, biomarkers, and therapies for these diseases, which substantially impact the US population.

Methods

Sex as a biological variable. Our study examined sex as a biological variable. We included cis-females (referred to as females) and cis-males (referred to as males) with no age restriction, according to gender information provided in the EHR.

Data sources. In this observational study, we obtained EHR data from January 1, 2011, to June 1, 2022 from the University of Southern California Health System (USC), a large multispecialty health system

with 2 inpatient tertiary care centers and multiple outpatient specialty clinics across the Los Angeles area; the University of Florida and Shands Health System (UF/Shands), an academic medical network with 11 hospitals and numerous outpatient clinics located in Florida; Mass General Brigham Health System (MGB), a Boston-based nonprofit hospital and physician network; University of Iowa Health Care, the only academic health system in the state that is centrally located in Iowa City, which used Iowa Health Data Resource (30); Medical College of Wisconsin (MCW), a private academic medical center with extensive clinical partnerships across Wisconsin; and Washington University School of Medicine in St. Louis (WUSTL), a private research university partnered with Barnes-Jewish Hospital (Figure 1).

Study population. For the denominator, we included patients that had at least 2 diagnoses of any disease at least 30 days apart (the denominator algorithm) (Figure 1). For the numerator, a patient was determined to have a diagnosis of an autoimmune disease if they had at least 2 diagnoses codes for the disease at least 30 days apart (the numerator algorithm). We examined EHR records collected between January 1, 2011 and June 1, 2022. We describe considerations for this analysis strategy below and acknowledge that different approaches affect prevalence outcomes. We want to emphasize, however, that a goal of this manuscript was to provide a program that is freely available for clinicians and researchers to use their own strategies and datasets to arrive at overall and individual US autoimmune disease prevalence estimates.

To test the accuracy of our algorithm, we conducted sensitivity analyses to determine how changing the number of diagnosis codes and the number of days between diagnosis codes (the date window) affected both the numerators and denominators used in our prevalence estimate (Supplemental Tables 5–7). Because the EHR is used for billing purposes in the US, a patient may receive a provisional diagnosis of an autoimmune disease to justify ordering tests to rule out the disease (31). While provisional diagnoses are also used in other countries, they are not required for billing purposes, whereas the US medical system makes such diagnoses a financial requirement (32). Thus, the use of a single diagnostic code to classify a patient as diagnosed with a disease will likely be an inaccurate source for determining prevalence. Supplemental Table 5 demonstrates that use of a single diagnosis code (0 date window) would overstate case counts by 31%–53% (average 41%) if 6 diseases were analyzed.

To investigate the effect of changing the denominator we found that using 2 diagnosis codes and a 30-day window gave a prevalence estimate of 5.9%, while other date windows ranging from 60–720 produced prevalence estimates of around 6.1%–6.5% (Supplemental Table 6). Thus, the prevalence gets larger as the denominator gets smaller with larger date windows because the algorithm catches fewer people. The percent change in prevalence by altering the denominator from 0–30 days or more was around 18% (Supplemental Table 7). However, the prevalence calculated using 2 codes over increasing date windows varied only by a small percentage, indicating that a 30-day date window was a valid and conservative estimate of prevalence (Supplemental Table 7). Based on these analyses, we required 2 diagnostic codes over a minimum time period (the date-window) of 30 days to classify patients as being diagnosed with a specific autoimmune disease. A study by Chung et al. (22) in 2013 also found that the use of 2 diagnostic codes provided improved specificity when using EHR data to identify patients diagnosed with RA. A code of "Autoimmune disease not otherwise classified" plus a specific disease code recorded 30 or more days later also qualified a patient as being diagnosed with a specific autoimmune disease.

To further validate the denominator, and the algorithm generally, we implemented an algorithm for RA developed by researchers at Harvard Medical School for use on the Electronic Medical Records and Genomics (eMERGE) network, a national network organized and funded by the National Human Genome Research Institute. The algorithm, posted on the Phenotype KnowledgeBase (PheKB) as Phenotype 585, is a machine-learning logistic regression model that uses a combination of log-weighted factors to classify patients with and without RA (https://phekb.org/phenotype/rheumatoid-arthritis-ra). The area under the receiver operating curve (AUROC) for the algorithm is 0.95 (see previous URL).

When we ran the eMERGE algorithm on the USC dataset, the program classified 2,552 patients with RA. Our denominator algorithm (2 diagnosis codes at least 30 days apart) computed the USC denominator as 375,253 for a prevalence of 6.8% (Supplemental Table 8). Extending to the US population gave an estimated prevalence of 2,264,648, which is within a published estimated prevalence for RA of 1,099,890 to 2,633,070 (33). Using our algorithm and sex- and age-adjusted data from USC estimates a prevalence of 2,586,344 individuals or 7.8%. Our algorithm across all 6 sites for RA estimates a prevalence of 2,580,060 individuals or 7.7% (Supplemental Table 8).

The date of death is not well tracked in electronic medical records. For sites that provided these data, the algorithm removed the patients. However, at sites without the date of death, patients remain in both the numerator and denominator, so death does not materially alter the prevalence estimate.

Selection of autoimmune diseases. The list of 105 autoimmune diseases included in this study was based on the textbook, "The Autoimmune Diseases" by Rose and MacKay, 5th Edition (1), with addition of select autoimmune diseases to establish a list of diseases for which substantive published evidence exists (Supplemental Table 1). Our list included all autoimmune diseases for which there is evidence in the literature that self-reactive T cells and/or antibodies contribute to or cause the disease (22). A list of diseases that lack this evidence, but are often categorized as autoimmune, is included in Supplemental Table 2, along with published and computed prevalence, for thoroughness; however, these conditions were not included in our autoimmune disease prevalence estimates.

Statistics. The data were transformed into the Observational Medical Outcomes Partnership (OMOP) model by each institution's local information technology personnel, and ICD-9 and ICD-10 codes were transformed to the Systematized Nomenclature of Medicine (SNOMED) coding system.

Since our goal was to assess the prevalence of all autoimmune diseases in the US using a standardized and replicable methodology, we sought an algorithm for computing the numerators that met the following criteria: (a) the algorithm is applicable across all autoimmune diseases without being more selective for some diseases than others; (b) the algorithm can operate at many health systems, not just those with a specific EHR system; (c) the algorithm can be run repeatedly so that changes in statistics can be tracked longitudinally; and (d) the algorithm can serve as a basis for more complete algorithms in the future (for example, algorithms that include medications and lab tests in the EHR, as well as notes).

Projecting the site-based prevalence estimates from individual sites to the US population required a denominator for the 6 sites' populations. Computing a denominator using EHR data has challenges. In the US,

EHR data are siloed by healthcare organizations, but patients can cross from one organization to another for their care, especially for emergency visits. Healthy patients may not seek care at all, or when they do, they may go to consumer-oriented facilities outside of the healthcare organizations (e.g., CVS Minute Clinics). Finally, females use the health system more than males (34), which may bias the dataset because females are known to be affected by autoimmune disease more often than males.

To project age- and sex-adjusted prevalence, we stratified the numerator and denominator into 4 age groups across 2 sex categories. Numerator and denominators for each site were used to compute an age-sex ratio, and that ratio was applied to the corresponding age-sex population based on US Census Data for 2022. We then combined the prevalence projections for all 8 age and sex categories into a total projected prevalence for all diseases for each site (disease specific age- and sex-adjusted values have also been computed for validations that appear elsewhere in the paper).

The software program, made freely available to the research community and included in Supplementary Materials, allows the date-windows for the numerator and denominator to be changed independently to allow refinement of this analysis by other investigators.

Study approval. Retrospective review of the demographic and clinical data from the EHR reported in this manuscript was approved by the Institutional Review Board (IRB) of each site. The need for written informed consent was waived by each IRB. Initial development of the algorithms was performed at the University of Southern California under IRB HS-20-00902. The IRB at Mass General Brigham determined on 7/11/2023 that the use of deidentified data made the project nonhuman-subject research and the need for an IRB was waived (REDCap ID #691). Work at the University of Florida/Shands Health System was conducted under IRB 202201755. The IRB at University of Iowa determined the project (IRB# 202403419) was not human subject research on 03/21/24. The Medical College of Wisconsin IRB reviewed the study (ID PRO00051359) on 6/11/2024 and determined it did not meet criteria for human subject research. The IRB at Washington University in St. Louis' Research Data Core Repository determined that the project (IRB #201607071) reported only summary statistics and did not constitute human subject research. The research conformed to the principles outlined in the Declaration of Helsinki.

Data availability. The program code used to generate the data for the manuscript is included in Supplemental Material and is made freely available to the research community provided they acknowledge the manuscript source. The code is modifiable for future studies. All data generated in the study are provided in the manuscript and Supplemental Data Values files. All questions regarding the study and program code should be directed to the cosenior authors of the study.

Author contributions

AHA conceived the project and directed the study with input from all authors. AHA, MGW, and NRR were involved in study design. NB provided study oversight. MA, XH, JSZ, and CDH provided data access. AHA conducted data analysis. AHA and DF wrote the manuscript. AHA, IH, NB, SC, ABC, LMW, GPL, XH, SNM, CDH, JA, MGW, JSZ, EMA, MA, PROP, AML, HAD, AAH, CEO, ADG, BWT, KIO, DND, NRR, FWM, GCT, and DF were involved in data interpretation and editing the manuscript.

Acknowledgments

Support for the project was provided by the University of Southern California, University of Florida, Mass General Brigham, and the Autoimmune Registry Inc., a 501(c) (3) nonprofit organization. Research reported in this publication was supported by the National Institutes of Health (NIH) National Center for Advancing Translational Sciences under Award Number UM1TR004403 to HAD, TL1 TR002380 to DND and DF, and National Heart, Lung and Blood Institute grant R01 HL164520 to DF. This research was also supported in part by the Intramural Research Program of the NIH National Institute of Environmental Health Sciences to FWM. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We especially acknowledge the late Dr. Noel R. Rose, MD, PhD, who guided the initial research, and whose dedication to scientific inquiry inspired this work.

Address correspondence to: DeLisa Fairweather, PhD; Department of Cardiovascular Medicine, Mayo Clinic, 4500 San Pablo Road, Jacksonville, Florida 32224, USA. Phone: 904.953.6351; Email: Fairweather.DeLisa@mayo.edu.

1. Mackay IR, Rose NR, eds. *The Autoimmune Diseases. Fifth Edition*. Elsevier; 2013.
2. Jacobson DL, et al. Epidemiology and estimated population burden of selected autoimmune diseases in the United States. *Clin Immunol Immunopathol*. 1997;84(3):223–243.
3. Eaton WW, et al. The prevalence of 30 ICD-10 autoimmune diseases in Denmark. *Immunol Res*. 2010;47(1–3):228–231.
4. Cooper GS, et al. Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. *J Autoimmun*. 2009;33(3-4):197–207.
5. Hayter SM, Cook MC. Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. *Autoimmun Rev*. 2012;11(10):754–765.
6. Conrad N, et al. Incidence, prevalence, and co-occurrence of autoimmune disorders over time and by age, sex, and socioeconomic status: a population-based cohort study of 22 million individuals in the UK. *Lancet*. 2023;401(10391):1878–1890.
7. Miller FW. The increasing prevalence of autoimmunity and autoimmune diseases: an urgent call to action for improved understanding, diagnosis, treatment, and prevention. *Curr Opin Immunol*. 2023;80:102266.
8. National Cancer Institute. SEER Cancer Statistics Review, 1975-2018. https://seer.cancer.gov/csr/1975_2018/. Updated April 15, 2021. Accessed December 12, 2024.
9. Center for Disease Control. Heart Disease Facts. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html. Accessed December 12, 2024.
10. Hay RJ, et al. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. *J Invest Dermatol*. 2014;134(6):1527–1534.
11. GBD Chronic Respiratory Disease Collaborators. Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Respir Med*. 2020;8(6):585–596.
12. Mathews SC, et al. Prevalence and financial burden of digestive diseases in a commercially insured population. *Clin Gastroenterol Hepatol*. 2022;20(7):1480–1487.
13. Westbury CF. Bayes' rule for clinicians: an introduction. *Front Psychol*. 2010;1:192.
14. Cooper GS, et al, eds. *Enhancing NIH Research on Autoimmune Disease*. The National Academies Press; 2022.
15. Tsao CW, et al. Heart disease and stroke statistics-2023 update: a report from the American Heart Association. *Circulation*. 2023;147(8):e93–e621.
16. Sardu C, et al. Population based study of 12 autoimmune diseases in Sardinia, Italy: prevalence and comorbidity. *PLoS One*. 2012;7(3):e32487.
17. Eaton WW, et al. Comorbidity of autoimmune diseases: a visual presentation. *Autoimmun Rev*. 2020;19(10):102638.
18. Farh KK, et al. Genetic and epigenetic fine

mapping of causal autoimmune disease variants. *Nature*. 2015;518(7539):337–343.
19. Dinse GE, et al. Increasing prevalence of antinuclear antibodies in the United States. *Arthritis Rheumatol*. 2020;72(6):1026–1035.
20. Health Central. Have you Noticed? Autoimmune Diseases are on the rise. https://www.healthcentral.com/article/rise-in-autoimmune-diseases. Updated October 22, 2020. Accessed December 12, 2024.
21. Autoimmune Association. Autoimmunity: Immune system fighting power. https://autoimmune.org/autoimmunity-immune-system-fighting-power/. Accessed December 12, 2024.
22. Chung CP, et al. A systematic review of validated methods for identifying patients with rheumatoid arthritis using administrative or claims data. *Vaccine*. 2013;31 Suppl 10:K41–K61.
23. United States Census Bureau. U.S. and World Population Clock. https://www.census.gov/popclock/. Updated December 12, 2024. Accessed December 12, 2024.
24. Castro C, Gourley M. Diagnostic testing and interpretation of tests for autoimmunity. *J Allergy Clin Immunol*. 2010;125(2 suppl 2):S238–S247.
25. Solomon AJ, et al. Misdiagnosis of multiple sclerosis: Impact of the 2017 McDonald criteria on clinical practice. *Neurology*. 2019;92(1):26–33.
26. Broers MC, et al. Misdiagnosis and diagnostic pitfalls of chronic inflammatory demyelinating polyradiculoneuropathy. *Eur J Neurol*. 2021;28(6):2065–2073.
27. Gardner TB, et al. Misdiagnosis of autoimmune pancreatitis: a caution to clinicians. *Am J Gastroenterol*. 2009;104(7):1620–1623.
28. Welters A, et al. Characterization of diabetes following pancreatic surgery in patients with congenital hyperinsulinism. *Orphanet J Rare Dis*. 2018;13(1):230.
29. National Institutes of Health. ICD Coding for Rare Diseases. https://rarediseases.info.nih.gov/tips/pages/123/icd-coding-for-rare-diseases. Accessed December 12, 2024.
30. Davis HA, et al. The Iowa Health Data Resource (IHDR): an innovative framework for transforming the clinical health data ecosystem. *J Am Med Inform Assoc*. 2024;31(3):720–726.
31. Bastarache L, et al. Developing real-world evidence from real-world data: transforming raw data into analytical datasets. *Learn Health Syst*. 2022;6(1):e10293.
32. Centers for Medicare and Medicaid Services. Medicare Claims Processing Manual: Chapter 16 - Laboratory Services. https://www.cms.gov/regulations-and-guidance/guidance/manuals/internet-only-manuals-ioms-items/cms018912. Accessed December 12, 2024.
33. Tuncer T, et al. Prevalence of rheumatoid arthritis and spondyloarthritis in Turkey: A nationwide study. *Arch Rheumatol*. 2018;33(2):128–136.
34. Bertakis KD, et al. Gender differences in the utilization of health care services. *J Fam Pract*. 2000;49(2):147–152.
35. United States Census Bureau. US Census: 1996, 2006, 2012, and 2022. https://www.census.gov/. Accessed December 12, 2024.