

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

1 **TITLE PAGE**

2

3 **Title:**

4 **Performance of a Computational Phenotyping Algorithm for Sarcoidosis Using**

5 **Diagnostic Codes in Electronic Medical Records: A Pilot Study from Two Veterans**

6 **Affairs Medical Centers**

7

8 **Authors:**

9 **Mohamed I Seedahmed, MD, MPH^{1,2*}, Izabella Mogilnicka, MD^{1,5*}, Siyang Zeng, MS^{1,4},**

10 **Gang Luo, PhD⁴, Charles McCulloch, PhD³, Laura Koth, MD²^γ, Mehrdad Arjomandi,**

11 **MD^{1,2}^γ**

12 ¹ San Francisco Veterans Affairs Medical Center, San Francisco, California, USA

13 ² Division of Pulmonary, Allergy, Critical Care and Sleep Medicine, Department of
14 Medicine, University of California San Francisco, California, USA

15 ³ Department of Epidemiology & Biostatistics, University of California San Francisco,
16 California, USA

17 ⁴ Department of Biomedical Informatics and Medical Education, School of Medicine,
18 University of Washington, Seattle, Washington, USA

19 ⁵ Department of Experimental Physiology and Pathophysiology, Laboratory of the Centre for
20 Preclinical Research, Medical University of Warsaw, Warsaw, Poland

21

22 * MIS and IM: These authors contributed equally to this work.

23 ^γMA and LK: These authors share senior authorship.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

24 **Email Addresses | ORCID iDs:**

25 MIS: mohamed.seedahmed@ucsf.edu | <https://orcid.org/0000-0002-7446-7346>

26 IM: izabella.mogilnicka@gmail.com | <https://orcid.org/0000-0002-5735-8375>

27 SZ: siyang.zeng@ucsf.edu | <https://orcid.org/0000-0001-9346-301X>

28 GL: luogang@uw.edu | <https://orcid.org/0000-0001-7217-4008>

29 CM: charles.mcculloch@ucsf.edu | <https://orcid.org/0000-0002-1279-6179>

30 LK: laura.koth@ucsf.edu | <https://orcid.org/0000-0001-9541-3622>

31 MA: mehrdad.arjomandi@ucsf.edu | <https://orcid.org/0000-0002-0116-9217>

32

33 **Word Count:** 3,829 words.

34

35 **Corresponding Author:**

36 **Mohamed I Seedahmed, MD, MPH**

37 Division of Pulmonary, Critical care, allergy and Immunology, and Sleep.

38 Department of Medicine, University of California, San Francisco

39 University of California San Francisco Helen Diller Medical Center

40 513 Parnassus Ave.

41 HSE 1314, Box 0111

42 San Francisco, CA 94143

43 Office: (415) 476-0735

44 Fax: (415) 502-2605

45 EMAIL: mohamed.seedahmed@ucsf.edu

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

46 **Authors' Contributions:**

47 All authors read and approved the final manuscript.

48 Conceived and designed the study research: MIS, LK, MA

49 Developed study protocol: MIS, LK, MA

50 Worked on the methods: MIS, IM, SZ, CM, LK, MA

51 Analyzed and Interpreted data: MIS, IM, SZ, CM, LK, MA

52 Prepared and/or edit the manuscript: MIS, IM, GL, CM, LK, MA

53

54 **Keywords:**

55 Sarcoidosis, Electronic Medical Records (EMR), Computational Phenotyping, Diagnostic

56 Codes, Veterans Affairs (VA), ATS Practice Guidelines.

57

58 **Highlights**

59 • Identifying sarcoidosis cases using diagnostic codes in EMR has low accuracy.

60 • “Unstructured data” contain information useful in identifying cases of sarcoidosis.

61 • Computational algorithms could improve the accuracy and efficiency of

62 case identification in EMR.

63 • We introduce a new scoring system for assessing healthcare providers' compliance with

64 the American Thoracic Society (ATS) practice guideline.

65 • Compliance scoring could help automatically assess sarcoidosis patients' care delivery.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

66 **ABSTRACT (250 words)**

67 **Background-**The accuracy of identifying sarcoidosis cases in electronic medical records
68 (EMR) using diagnostic codes is unknown.

69 **Methods-**To estimate the statistical performance of using ICD-9 and ICD-10 diagnostic
70 codes in identifying sarcoidosis cases in EMR, we searched the San Francisco and Palo Alto
71 Veterans Affairs (VA) medical centers EMR and randomly selected 200 patients coded as
72 sarcoidosis. To further improve diagnostic accuracy, we developed an “index of suspicion”
73 algorithm to identify probable sarcoidosis cases based on clinical and radiographic features.
74 We then determined the positive predictive value (PPV) of diagnosing sarcoidosis by two
75 computational methods using ICD only and ICD plus the “index of suspicion” against the
76 gold standard developed through manual chart review based on the American Thoracic
77 Society (ATS) practice guideline. Finally, we determined healthcare providers’ adherence to
78 the guidelines using a new scoring system.

79 **Results-**The PPV of identifying sarcoidosis cases in VA EMR using ICD codes only was 71%
80 (95%CI=64.7%-77.3%). The inclusion of our construct of “index of suspicion” along with the
81 ICD codes significantly increased the PPV to 90% (95%CI=85.2%-94.6%). The care of
82 sarcoidosis patients was more likely to be classified as “Fully” or “Substantially” adherent with
83 the ATS practice guideline if their managing provider was a specialist (45% of primary care
84 providers vs. 74% of specialists; P=0.008).

85 **Conclusions-**Although ICD codes can be used as reasonable classifiers to identify
86 sarcoidosis cases within EMR, using computational algorithms to extract clinical and
87 radiographic information (“index of suspicion”) from unstructured data could significantly
88 improve case identification accuracy.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

89 INTRODUCTION

90 Sarcoidosis is a complex disease of unknown etiology that can involve multiple organs,
91 with no universal and standardized measures that fully secure the final diagnosis.[1–3] In fact,
92 only recently the American Thoracic Society (ATS) published its first practice guideline to
93 provide recommendations for diagnosing sarcoidosis and the necessary screening tests.[3] The
94 ATS practice guideline for diagnosis requires the presence of specific clinical and radiographic
95 features, tissue biopsy revealing non-necrotizing granulomas, and exclusion of alternative
96 conditions that can mimic sarcoidosis.[1,3,4]

97 Electronic medical record (EMR) data are commonly used in research and by healthcare
98 systems, including the Department of Veterans Affairs, to predict outcomes or assess care
99 quality.[5] EMR data are generally captured in two forms: 1) *structured data* such as billing
100 codes like International Classification of Disease (ICD) codes, laboratory test results,
101 procedural codes, and 2) *narrative or unstructured data* such as progress notes, pathology
102 reports, and imaging reports. Unstructured data contain many more details of the clinical
103 conditions, but extracting these details is challenging and time-consuming. In contrast,
104 structured data are easier to search and offer the promise to identify cases computationally
105 using diagnostic codes. But, diagnostic codes can be inaccurate and not easily verifiable,
106 particularly for the case definition of complex diseases such as sarcoidosis.[6–8] Thus, there is
107 a need to develop automated algorithms to verify the final diagnosis of sarcoidosis using
108 unstructured data by incorporating the ATS diagnostic criteria.

109 The goals of this study are twofold. First, we sought to estimate the statistical
110 performance of using diagnostic codes (ICD-9 and 10) to identify patients with sarcoidosis in
111 the United States Veterans Affairs (VA) EMR through the VA Informatics and Computational
112 Infrastructure (VINCI) via applying a recently published ATS practice guideline. Second, we

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

113 investigate healthcare providers' practice patterns and determine how adherent these patterns
114 are with the new practice guideline issued by the ATS. This study will help researchers and
115 healthcare systems understand the healthcare providers' compliance with these new
116 recommendations and conduct research using administrative databases.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

117 **METHODS**

118 **Study Design**

119 We determined the statistical performance of using diagnostic codes (the International
120 Classification of Diseases, Ninth and Tenth Revisions codes; ICD-9 and ICD-10) to identify
121 patients with sarcoidosis from the EMR. First, by identifying patients with ICD diagnosis code
122 of sarcoidosis, and then performing a comprehensive chart review of the entire unstructured
123 data to ascertain the true diagnosis of sarcoidosis. Furthermore, to improve the diagnostic
124 accuracy using ICD code alone, we generated an “index of suspicion” criteria from
125 unstructured clinical and radiographic data, but not histopathologic data, as a second decision
126 point along with ICD code. We then determined the positive predictive value of both the use
127 of ICD code and ICD code plus “index of suspicion” to determine whether the use of “index
128 of suspicion” could improve the accuracy of identifying sarcoidosis cases in EMR regardless
129 of tissue biopsy availability. Finally, we investigated the relevant healthcare utilization among
130 the identified patients to determine healthcare providers’ adherence to the ATS practice
131 guideline.

132

133 **Data Source**

134 This is a retrospective cohort study of electronic medical records (EMR) available from
135 the Veterans Affairs Informatics and Computational Infrastructure (VINCI). Developed by the
136 VA Health Services Research & Development (HSR&D) to improve veterans’ healthcare,
137 VINCI provides access to comprehensive and integrated veterans’ national datasets that are de-
138 identified and the necessary computational and analytical tools in a secure, high-performance
139 computing environment [9]. The University of California San Francisco Institutional Review
140 Board and the Veterans Health Administration Research and Development Committee
141 approved this study.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

142 **Data Collection**

143 We searched the EMR data in VINCI from 1989 to 2019 and identified all patients
144 coded as sarcoidosis cases in the VA healthcare system, as defined by documentation of the
145 ICD-9 and 10 codes of 135 and D.86 (including all subcodes), respectively. Data were
146 extracted through executing Structured Query Language (SQL) queries in a SQL Server 2017
147 database.

148

149 We then performed a comprehensive and thorough manual chart review of the
150 unstructured data (available clinical notes, pathology reports, radiology reports, and laboratory
151 test results) to determine the accuracy of the ICD-9 or 10 code-based diagnosis of sarcoidosis
152 among cases. To limit the required chart review to a manageable level, we reviewed a total of
153 200 cases out of the 14,833 identified. Because our access to the detailed medical records was
154 limited to two VA medical centers (San Francisco VA [SFVA] and Palo Alto Medical Centers
155 [PAVA]), the 200 patients coded as sarcoidosis were selected from those two centers. We
156 stratified the list of sarcoidosis cases from the two centers by site and used the “lottery” method
157 to select 100 patients from each site randomly.[10]

158 Two independent reviewers (MIS, IM) confirmed the diagnoses of sarcoidosis by
159 performing a manual chart review of the 200 cases in the VA EMR, Computerized Patient
160 Record System (CPRS), and based on the ATS diagnostic criteria (clinical, radiographic,
161 pathological findings, and exclusion of other causes).[3] This approach was considered the
162 “gold standard” methodology for diagnostic accuracy.

163 Given that in clinical practice, not all suspected patients will have a tissue biopsy, we
164 generated an “index of suspicion” for sarcoidosis to identify patients with probable sarcoidosis
165 based on clinical and radiographic information and regardless of the availability of biopsy data.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

166 The “index of suspicion” was defined based on the interpretation of available unstructured data
167 from both clinical and radiographic findings without the inclusion of tissue biopsy results if
168 available and was used to determine the likelihood of an individual having sarcoidosis. Upon
169 completing the chart review, each patient was assigned to one of two groups of “high index of
170 suspicion” if the clinical and radiological presentations were supportive of sarcoidosis, or “low
171 index of suspicion” if clinical and radiological findings were not consistent with sarcoidosis.
172 We then classified the patients into three groups: (1) Patients with a high index of suspicion
173 and documented histopathological evidence of non-necrotizing granulomas were categorized
174 as “probable sarcoidosis with confirmed biopsy”; (2) Those with a high index of suspicion and
175 either no documented biopsy in the EMR or a biopsy showing no histopathological evidence
176 of non-necrotizing granulomas were categorized as “probable sarcoidosis without confirmed
177 biopsy”; and (3) those with a low index of suspicion were classified as “unlikely sarcoidosis.”

178 This approach was used to compare the statistical performance of the two methods (ICD
179 code alone versus ICD code with “index of suspicion”) in identifying sarcoidosis patients from
180 the EMR. As we started with a random sample of those with sarcoidosis diagnostic codes (ICD
181 codes 135 or D86), the further restriction of the sample to those with “high index of suspicion”
182 is still a random sample of the combination of both ICD codes and “index of suspicion.”

183 **Disease-related Variables**

184 Organ involvement was assessed based on the clinical history obtained from
185 physicians’ notes and imaging and biopsy reports available in CPRS. For this assessment, to
186 adjust for the variability in providers’ documentation, we adapted a set of criteria previously
187 introduced in the NIH-sponsored Genomic Research in Alpha-1 Antitrypsin Deficiency and
188 Sarcoidosis (GRADS) Study.[11]

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

189 We collected the following data from chart review: clinical site (SFVA or PAVA),
190 gender, race, ICD-9 and 10 code for sarcoidosis (135 and D86), index date of ICD code (entry
191 date of the ICD code), specialty visits (specialty notes available), the pathological diagnosis
192 from any available biopsy, organ involvement (as described in **Figure 2**), Scadding staging of
193 chest x-ray (as described in radiology reports), history of bilateral hilar lymphadenopathy
194 (based on radiology reports and clinical notes), pulmonary function test (**PFT**) pattern (as
195 reported in PFT reports), vitamin D levels (25 [OH]D and 1,25[OH]D), and finally the
196 treatment status of sarcoidosis. Besides, we checked whether other imaging and laboratory
197 reports were performed and available in EMR as part of diagnostic workup and sarcoidosis
198 management; including echocardiography, 12-lead electrocardiogram (**ECG**), chest x-ray,
199 chest computed tomography (**CT**), cardiac magnetic resonance imaging (**MRI**) or positron
200 emission tomography (**PET**)/ CT, or abdominal CT scan or ultrasound.

201 Pathological diagnoses were categorized into “primary” histopathological if it was
202 available in the pathology report domains, or “secondary” if the data was only available in
203 clinical note domains due to either a remote history of biopsy or the biopsy being performed
204 outside of the VA. The PFT reports at SFVA and PAVA used Crapo reference equations to
205 calculate the Lower Limit of Normal (**LLN**) values for spirometry and lung volume
206 measurements.

207 Using the clinical data from chart abstraction, we classified the patients into the clinical
208 phenotypes proposed by the GRADS study with some modifications.[11] Because of the
209 variability of available data needed to confirm multiple organ involvement from EMR chart
210 review, we considered three or more organ involvement (instead of five or more as defined in
211 the GRADS study) to categorize “multiorgan phenotype.”

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

212 Finally, we created a scoring system to estimate the physicians' compliance with the
213 recently published ATS clinical practice guideline.[3] For this analysis, four variables were
214 used in the scoring system: (1) having an ophthalmology visit, (2) 12-lead EKG, (3) chest X-
215 ray, and (4) vitamin D measurements (25-(OH) D and 1,25(OH) D) within one year after the
216 index date of ICD code. Each variable was given a score of three, and each patient could have
217 a total score of 12. We classified the physicians' compliance into four categories based on the
218 score: "Fully," "Substantially," "Partially," or "Non-compliant" for scores of 12, 9-11, 5-8, or
219 <5 points, respectively. This scoring system was applied to all probable sarcoidosis patients
220 with or without histopathology confirmation. Furthermore, the role of primary versus specialty
221 care evaluation and management was examined by comparing the healthcare providers'
222 compliance score.

223 **Statistical Analyses**

224 All statistical analyses were performed with R software (RStudio, version 1.2.5,
225 Foundation for Statistical Computing, Vienna, Austria). Descriptive statistics were computed
226 to summarize the data. Categorical variables were presented as the frequency in percentages
227 and continuous data as means and standard deviations. We estimated the positive predictive
228 value (PPV) of the two computational diagnostic criteria for sarcoidosis described above (ICD
229 codes only and ICD codes along with "index of suspicion"). The PPV of the criterion of using
230 ICD code only was calculated as the number of patients verified to have sarcoidosis by chart
231 review ("gold standard" or true positives) divided by the total number of patients with an ICD-
232 9 or 10 diagnostic code of sarcoidosis. The PPV of the criterion of using ICD codes and "index
233 of suspicion" use was calculated as the number of patients verified to have sarcoidosis ("gold
234 standard" or true positives) divided by the total number of patients with a high index of
235 suspicion. We computed the 95% confidence intervals (CI) using the exact binomial method.
236 For our estimates, the significance was defined as $P < 0.05$.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

237 To determine whether it is worth interpreting the developed contingency table that
238 compares the healthcare providers' compliance score by primary versus specialty care, we
239 computed the chi-squared test of significance.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

240 **RESULTS**

241 **Patients' Characteristics**

242 A total of 14,833 patients with at least one ICD-9 or 10 diagnostic code of sarcoidosis
243 were identified (**Figure 1**). The study cohort included 200 patients identified by ICD codes of
244 sarcoidosis: 169 with ICD-9 code of 135 and 39 with ICD-10 code of D.86 (**Table 1**). Of the
245 200 patients, 158 had a “high index of suspicion” for sarcoidosis based on clinical and
246 radiographic findings. Of those, 142 had confirmed sarcoidosis based on histopathological
247 evidence of non-necrotizing granuloma and were classified as “probable sarcoidosis with
248 confirmed biopsy.” The remaining 16 patients with a “high index of suspicion” did not undergo
249 a biopsy and were classified as “probable sarcoidosis without confirmed biopsy” (**Figure 2**).
250 No patient had nondiagnostic biopsy results for sarcoidosis.

251 Among those patients with probable sarcoidosis (with and without confirmed biopsy),
252 90% (142/158) were males (**Table 1**), and there was a higher representation of African
253 Americans than non-Hispanic Whites (54% (85/158) vs. 33% (52/158), respectively). Overall,
254 90% (143/158) had a predominant pulmonary phenotype, and 29% (45/158) had a multi-organ
255 disease that included pulmonary. Among those with pulmonary phenotype, 28% (36/129), 22%
256 (28/129), and 19% (25/129) had restrictive, obstructive, and mixed lung function patterns,
257 respectively. The majority were in Scadding stage II (32%, 47/145), followed by Stage 0 and
258 Stage 1 (20% (29/145) and 18% (26/145), respectively). There was no significant difference in
259 age between those who did and those who did not have a biopsy performed to diagnose
260 sarcoidosis (mean age= 65.5 vs. 69.3, P= 0.18, respectively). In terms of clinical phenotypes,
261 the most common phenotype was multiorgan (33%, 52/158), followed by stage II or III treated
262 (29%, 45/158). Our study cohort did not include any individuals with acute presentation (acute,
263 untreated). Some patients with remitting phenotype overlapped with Groups 2 and 4 (**Table 1**)
264 (**Figure 3**).

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

265 **Diagnostic Accuracy of ICD Codes**

266 We then calculated the positive predictive value (PPV) of using ICD codes to identify
267 VA patients that met the ATS definition of sarcoidosis from the VINCI database. For this
268 calculation, we used the curated dataset of 200 patients. The estimated PPV of using only ICD
269 codes was 71% (95%CI= 64.7%-77.3%). The inclusion of our construct of “index of suspicion”
270 along with the ICD codes into the identification of sarcoidosis increased the PPV estimate
271 significantly to 90% (95%CI= 85.2%-94.6%) (**Table 2**), as the initial sarcoidosis cohort was
272 restricted to the patients with “high index of suspicion.”

273

274 **Providers’ Compliance with the ATS Clinical Practice Guideline**

275 Among those with probable sarcoidosis (with and without biopsy), 13% were managed
276 by primary care providers only, 51% (81/158) were managed by pulmonary physicians only,
277 and pulmonary physicians managed 36% (57/158) along with other specialties such as
278 ophthalmology or cardiology. The specialty care visits occurred in the context of diagnosis or
279 management of the disease.

280

281 Within one year of the entry date of the ICD code of sarcoidosis, 91% (143/158) of
282 patients had at least one chest x-ray, 83% (131/158) had at least one EKG, and 84% (133/158)
283 had at least one visit to an ophthalmologist. In addition, 53% (82/158) had 25(OH)D, and 23%
284 (36/158) had 1, 25(OH)D, including 20% (32/158) who had both measurements (**Table 1**).

285

286 Among the patients with probable sarcoidosis (with and without biopsy), 70%
287 (111/158) had managing providers who were “Fully” or “Substantially” compliant with the
288 ATS practice guideline based on our scoring system (**Table 3**). The majority of those patients,
289 92% (102/111), were managed by specialists, including pulmonary physicians, while primary

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

290 care providers managed 8% (9/111). Among the 30% (47/158) patients whose care only met
291 the “Partially compliant” or “Non-compliant” definition based on our scoring system, 23%
292 (11/47) were managed by primary care providers (**Figure 4**). This result indicates that the care
293 of sarcoidosis patients was more likely to be classified as “Fully” or “Substantially” adherent
294 with the ATS practice guideline if their managing provider was a specialist (45% [9/20] of
295 primary care providers vs. 74% [102/138] of specialists, $P = .008$) (**Table 4**).

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

296 **DISCUSSION**

297 In this retrospective cohort study of Department of Veterans Affairs EMR, we
298 evaluated 200 randomly selected patients diagnosed with sarcoidosis from the San Francisco
299 and Palo Alto Veterans Affairs Medical Centers. We found that using ICD-9 and 10 codes to
300 identify sarcoidosis patients within EMR has a relatively low accuracy (PPV of 71%
301 [95%CI= 64.7%-77.3%]) of detecting patients with “probable sarcoidosis” as defined by the
302 ATS clinical practice guideline. We also demonstrated that including an “index of suspicion”
303 that incorporates only the clinical and radiographical data allows for a significant increase in
304 diagnostic accuracy (PPV of 90% [85.2%-94.6%]), as the initial sarcoidosis cohort was
305 restricted to the patients with “high index of suspicion.” This “index of suspicion,” which
306 was developed based on the ATS clinical practice guideline [3], could be executed by manual
307 chart review of the available clinical and radiographic data and could potentially be adapted
308 for automated chart review algorithms using natural language processing and machine
309 learning in EMR. Furthermore, we found that the care provided by specialists was more
310 likely to be “Fully” or “Substantially” compliant with the ATS practice guideline compared
311 to primary care providers.

312

313 This randomly selected cohort of 200 Veterans with sarcoidosis diagnosis consisted of
314 90% men and 10% women. While the sex distribution in our study was different from the
315 ACCESS study[12], it is closely reflective of the veterans' population demographics.[13]
316 This study further confirmed the higher prevalence of sarcoidosis in African-Americans
317 (54%) compared to Caucasians (33%), which has been previously reported by many other
318 sarcoidosis epidemiology studies [14–18]. At the same time, the study population was
319 racially diverse, highlighting the potential utility of the VA EMR for studying sarcoidosis in
320 populations of people of color.[19]

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

321 The novelty of this study is the development of a scoring system to objectively measure
322 physicians' compliance with the latest ATS practice guideline. Adherence to recommendations
323 might allow for earlier interventions in sarcoidosis management and ultimately improve
324 outcomes by delaying or preventing the development of morbid phenotypes. Variables used in
325 this scoring system included ophthalmology visit, 12-lead EKG, chest x-ray, and vitamin D
326 levels measurements within one year of the entry date of the ICD code. These variables were
327 chosen based on their utility and importance.[3] Not all tests and measurements discussed in
328 the guidelines have been included in our scoring system. Yet, we believe it can be easily
329 modified to include other clinical variables. Notably, a longitudinal analysis of the association
330 between physicians' compliance with the recommendations and clinical outcomes in
331 sarcoidosis could provide direction for future investigations and ultimately help better guide
332 clinical practice to be used as a metric to assess the improvement of the care delivery for
333 sarcoidosis patients.

334

335 Using ICD codes alone to extract health information is far more convenient than the
336 time-consuming manual review process of narrative datasets in unstructured data.[8] However,
337 using ICD codes to identify sarcoidosis cases in large datasets with thousands of patients poses
338 several practical challenges. First, given the heterogeneity of sarcoidosis, it is challenging to
339 confirm the disease's presence efficiently. The verification process requires careful analysis of
340 the available narrative data such as progress notes, imaging reports, and pathology reports to
341 establish the case definition based on the sarcoidosis diagnostic criteria.[3] Second, the precise
342 identification of the type of organ involvement through EMR is a complex process and requires
343 a thorough review of unstructured data. Although there are sub-codes for ICD diagnostic codes
344 that aim to capture various organ involvement, healthcare providers may or may not be familiar
345 with those sub-codes and may or may not use them correctly. Moreover, there are no specific

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

346 ICD codes for classifying some of the various organ involvement in sarcoidosis (such as central
347 nervous system or GI tract).[20] Third, ICD codes do not determine the extent of disease, such
348 as that described by chest x-ray stages [21], due to a lack of ICD codes for different stages of
349 pulmonary sarcoidosis.[20] Analysis of pulmonary features requires a manual review of every
350 patient's radiology reports and cannot be performed using only ICD codes alone. Finally, ICD
351 codes do not specify the various sarcoidosis presentations such as acute, remitting, or chronic
352 disease.[11,20] Thus, they cannot be used to classify patients into previously described
353 phenotype groups.

354

355 The definition of clinical phenotypes has become an essential goal for the sarcoidosis
356 scientific community, as genetic studies have identified different patterns of gene expression
357 associated with disease severity and disease course.[22,23] In 2015, the National Heart, Lung,
358 and Blood Institute (NHLBI) held a workshop to leverage current scientific knowledge and
359 defining platforms to address disease disparities, identify high-risk phenotypes, and improve
360 sarcoidosis outcomes [24]. Nine different steps and research strategies were recommended to
361 expand the scope of sarcoidosis research, including EMR-based research to provide a unified,
362 multidisciplinary approach to bring together stakeholders interested in reducing the burden and
363 severity of sarcoidosis. However, the major barrier in the efficient use of EMR data is the
364 accurate extraction of research-quality variables, case definitions, and outcomes.[25] Thus, the
365 rapid identification of cases and extraction of relevant clinical variables from EMR using
366 computable phenotype algorithms have emerged as an important next step in EMR-based
367 research. Computable phenotype definitions are also essential to conducting pragmatic clinical
368 trials and comparative effectiveness research, increasing the healthcare system's capacity to
369 deliver Precision Medicine effectively.[26]

370

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

371 The two most commonly applied approaches to defining computable phenotypes are:
372 (1) a “*high-throughput*” phenotype algorithm using only structured data (traditionally, the ICD
373 diagnosis codes). (2) a “*low-throughput*” phenotype algorithm that accesses structured and
374 unstructured data to develop a sequential flow chart that should end with a case definition.
375 Such a low-throughput approach employs high-performance computational tools (such as
376 natural language processing [**NLP**]) to process text and extract information utilizing linguistic
377 rules instead of labor-intensive manual review by researchers.[6] This approach should
378 streamline the development of registries and help enrich EMR-based research studies.[27] Our
379 study highlights the need for the development of such automated methods to improve the
380 computational case-definition of sarcoidosis and other high-quality sarcoidosis-related
381 research variables, such as determining the date of the diagnosis, organ involvements, Scadding
382 stages, and the clinical status (acute, chronic, acute on the chronic or remitting disease).

383

384 Our study has several limitations. First, we used the ICD diagnosis code's entered date
385 as the surrogate time point to establish the patients' care with active sarcoidosis or re-
386 establish the care for those with remitting disease. This approach is different from what
387 sarcoidosis researchers traditionally use as a surrogate to determine the date of diagnosis,
388 which is the date of performing the biopsy. However, because the ICD diagnosis code's date
389 is more relevant for assessing the managing healthcare providers' compliance, we chose to
390 take the above approach. Second, in the cases where the biopsy report was unavailable (either
391 due to a remote history of biopsy or biopsy performed outside the VA), we relied on the
392 “secondary” histopathological reports documented in the providers' narrative within the
393 clinical notes. This approach made the diagnosis of sarcoidosis less robust because
394 confirmatory biopsy reports in those patients were not directly verifiable. However, we used
395 the index of suspicion approach to define probable sarcoidosis cases regardless of whether

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

396 there was a confirmatory biopsy report available, which is consistent with the diagnostic
397 algorithm recommended by the ATS practice guideline.[3] Last, the generalizability of our
398 findings in VA EMR to other populations could be limited because the Veterans form a
399 special population with different demographics and exposures from the general population.
400 However, the US Veterans Affairs Healthcare System EMR data cover over 22 million
401 Veterans across the US and over 14,000 patients with sarcoidosis ICD diagnosis codes,
402 providing an enormous number of patients to study a rare disease.

403

404 **Conclusion**

405 Although ICD codes can be used as reasonable classifiers to identify sarcoidosis cases
406 within EMR, using computational algorithms to extract clinical and radiographic information
407 (“index of suspicion”) from unstructured data could significantly improve case identification
408 accuracy. Using automated emerging methods (such as NLP) to develop a novel sarcoidosis-
409 specific computational phenotype algorithm could increase the efficiency of identifying these
410 cases from large healthcare databases. Furthermore, we found that specialists are more likely
411 than primary care providers to be “Fully” or “Substantially” compliant with the ATS clinical
412 practice guideline. This finding calls attention to the importance of examining the association
413 between physicians’ compliance with the recommendations and clinical outcomes
414 longitudinally.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

415 **TABLES:**

416 **Table 1.** Distribution of characteristics and clinical phenotype groups of patients with

417 probable sarcoidosis ^a (with and without confirmed biopsy)

	Probable Sarcoidosis ^a		
	Probable sarcoidosis with confirmed biopsy N (%)	Probable sarcoidosis without confirmed biopsy N (%)	p-value
Characteristics	142 (90)	16 (10)	
Age (mean years ±SD)	65.5 ±10.8	69.3 ±10.3	0.18
Gender			
M	127 (89.4)	15 (93.8)	0.59 ^c
F	15 (10.6)	1 (6.2)	
Race			
African American	74 (52.1)	11 (68.8)	0.62 ^b
Non-Hispanic White	49 (34.5)	3 (18.8)	
Hispanic White	3 (2.1)	0 (0)	
Unknown	12 (8.5)	2 (12.5)	
Other	4 (2.8)	0 (0)	
ICD codes for sarcoidosis			
ICD-9	98 (69)	10 (62.5)	0.60 ^c
ICD-10	44 (31)	6 (37.5)	
Specialty visit			
Pulmonary only	71 (50)	10 (63)	0.28 ^b
Other specialists (including multiple)	54 (38)	3 (18.5)	
Primary care only	17 (12)	3 (18.5)	

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

Ophthalmology visit	120 (84.5)	13 (81.2)	0.93 ^c
Organ involvement			
Lung	86 (60.5)	12 (75)	0.38 ^b
Multiorgan (pulmonary without cardiac involvement)	39 (27.5)	2 (12.5)	
Multiorgan (cardiac without pulmonary involvement)	2 (1.4)	0 (0)	
Multiorgan (not cardiac neither pulmonary involvement)	11 (7.8)	2 (0)	
Multiorgan (both cardiac and pulmonary involvement)	4 (2.8)	0 (0)	
EKG	118 (83.1)	13 (81.3)	0.97 ^c
PFT pattern ^d			
Obstructive	27 (19)	1 (6.2)	0.03 ^b
Restrictive	30 (21.1)	6 (37.5)	
Mixed	20 (14)	5 (31.3)	
Normal	39 (27.5)	1 (6.2)	
CXR	129 (90.8)	14 (87.5)	0.93 ^c
25(OH)D	72 (50.7)	10 (62.5)	0.80 ^c
1,25(OH)D	31 (21.8)	5 (31.3)	
Both (1,25 and 25 (OH)D)	31(21.8)	1(6.2)	

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

Scadding stage ^e			
Stage 0	24 (16.9)	5 (31.2)	0.06 ^b
Stage I	23 (16.2)	3 (18.8)	
Stage II	45 (31.7)	2 (12.5)	
Stage III	17 (12)	4 (25)	
Stage IV	22 (15.5)	0 (0)	
Clinical Phenotype Group ^f			
Group 1: Multiorgan	50 (35.2)	2 (12.5)	0.06 ^b
Group 2: Nonacute, Stage I, untreated	6 (4.2)	2 (12.5)	
Group 3: Stage II-III, treated	42 (29.6)	3 (18.8)	
Group 4: Stage II-III, untreated	14 (9.9)	2 (12.5)	
Group 5: Stage IV, treated	17 (12)	0 (0)	
Group 6: Stage IV, untreated	4 (2.8)	2 (12.5)	
Group 7: Acute sarcoidosis, untreated	0 (0)	0 (0)	
Group 8: Remitting, untreated	30 (21)	5 (31)	
Group 9: Cardiac sarcoidosis, treated	6 (4.2)	0 (0.0)	

418
419
420
421

^a Based on the clinical diagnostic criteria recommended by the American Thoracic Society (clinical, radiographic, and pathological findings) in the absence of other alternative diagnoses that could explain the presence of non-necrotizing granulomas on histopathology.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

422 ^b Fisher's exact test.

423 ^c X² test.

424 ^d Evaluated based on pulmonary function tests reports available in CPRS.

425 ^e Scored based on reviewers' interpretation of imaging reports using Scadding staging. Stage 0: normal chest
426 radiograph, Stage I: hilar or mediastinal nodal enlargement only; Stage II: nodal enlargement and parenchymal
427 disease; Stage III: parenchymal disease only; Stage IV: end-stage lung disease (pulmonary fibrosis).

428 ^g Clinical Phenotype Groups [11]:

429 Group 1: Multiorgan involvement: Patients with more than 2 organs involved.

430 Group 2: Nonacute, Stage I, untreated: Patients with nonacute sarcoidosis, Stage I, never treated for sarcoidosis.

431 Group 3: Stage II-III, treated: Patients with nonacute sarcoidosis, Stage II or III, formerly treated for sarcoidosis
432 or treated within 3 months of data review.

433 Group 4: Stage II-III, untreated: Patients with nonacute sarcoidosis, Stage II or III, never treated for sarcoidosis.

434 Group 5: Stage IV, treated: Patients with nonacute sarcoidosis, Stage IV, formerly treated for sarcoidosis or treated
435 within 3 months of data review.

436 Group 6: Stage IV, untreated: Patients with nonacute sarcoidosis, Stage IV, never treated for sarcoidosis.

437 Group 7: Acute sarcoidosis, untreated: Patients with acute sarcoidosis (Löfgren syndrome).

438 Group 8: Remitting, untreated: Patients who have had no evidence of active clinical disease for more than one
439 year.

440 Group 9: Cardiac sarcoidosis, treated: Patients with cardiac manifestations of sarcoidosis, formerly treated for
441 sarcoidosis or treated within 3 months of data review.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

442 **Table 2.** Contingency 2x2 table of using ICD and index of suspicion for sarcoidosis cases
443 identification.

	Index of Suspicion ^a	Sarcoidosis confirmed with a biopsy		
		Yes ^b	No ^c	Total
Patients with positive ICD 9 or ICD-10 for sarcoidosis	High index of suspicion	142 ^d	16 ^e	158
	Low index of suspicion	0 ^f	42 ^g	42
	Total	142	58	200

444

445 ^a Index of suspicion for sarcoidosis based on both clinical and radiographic evidence but not biopsy.

446 ^b Available biopsies with primary or secondary histopathological reports.

447 ^c No biopsies were ordered or available in the electronic medical record.

448 ^d Probable sarcoidosis group with histopathological evidence of non-necrotizing granuloma (NNG) = true
449 positives

450 ^e Probable sarcoidosis group without histopathological evidence of NNG= false positives

451 ^f No sarcoidosis group due to lack of sufficient clinical and radiological features consistent with sarcoidosis even
452 in the presence of the histopathological evidence of NNG = false negatives

453 ^g No sarcoidosis group due to lack of sufficient clinical and radiological features consistent with sarcoidosis, in
454 addition to the absence of the histopathological evidence of NNG= true negatives

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

455 **Table 3.** Healthcare providers' compliance with the ATS clinical practice guideline.

Healthcare providers' compliance with ATS practice guidelines ^a N (%)	Probable sarcoidosis with confirmed biopsy N=142	Probable sarcoidosis without confirmed biopsy N= 16
Fully compliant	26 (18.3)	1 (6.25)
Substantially compliant	76 (53.5)	8 (50)
Partially compliant	31 (21.9)	6 (37.5)
Non-compliant	9 (6.3)	1 (6.25)

456

457 ^a Developed with modifications based on recently published ATS practice guideline. The scoring system has
458 been developed based on the availability of the following within one year of the diagnosis: ophthalmology visit
459 = 3 points, 12-leads EKG = 3 points, CXR = 3 points, vitamin D (25-OH + 1,25-OH = 3 points; 25(OH)D or
460 1,25(OH)D = 2 points, otherwise zero points). Each variable weight a score of 3, with a total score of 12. Each
461 patient was given a final score, and the purpose was to classify the degree of the providers' compliance into four
462 categories based on the score:
463 o *Fully compliant*: if the score = 12
464 o *Substantially compliant*: if the score = 9-11
465 o *Partially compliant*: if the score = 5-8
466 o *Non-compliant*: if the score <5.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

467 **Table 4.** Contingency 2x2 table of the healthcare providers’ compliance score by primary
468 versus specialty care ^a.

Healthcare providers’ compliance with ATS practice guideline scored by our classification	Healthcare providers		
	Specialists	Primary care physicians	Total
“Fully” or “Substantially” compliant	102	9	111
“Partially” or “Non” compliant	36	11	47
Total	138	20	158

469

470 ^a X² statistic= 6.9877, P-value= 0.008.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

471 **Ethics approval and consent to participate:** The University of California San Francisco

472 Institutional Review Board and the Veterans Health Administration Research and

473 Development Committee approved this study. [IRB Protocol #15-16660].

474

475 **Availability of data and material:** Due to the sensitive nature of health data analyzed in the

476 current study, data will remain confidential and are not publicly available.

477

478 **Competing interests:** The authors have no conflicts of interest to disclose relevant to the

479 present work.

480 **Funding:** This work was supported by funds from the Department of Veterans Affairs

481 Fellowship Award to MIS; the Flight Attendants Medical Research Institute (FAMRI)

482 (CIA190001 to MA); Department of Veterans Affairs Clinical Sciences Research and

483 Development (CSRDP) (CXV-00125 to MA); the Tobacco-related Disease Research Program

484 of the University of California (T29IR0715 to MA).

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

485 **Bibliography**

- 486 [1] Judson MA. Advances in the diagnosis and treatment of sarcoidosis. *F1000Prime Rep*
487 2014;6:89. doi:10.12703/P6-89.
- 488 [2] Keller AZ. Anatomic sites, age attributes, and rates of sarcoidosis in U. S. veterans.
489 *Am Rev Respir Dis* 1973;107:615–20. doi:10.1164/arrd.1973.107.4.615.
- 490 [3] Crouser ED, Maier LA, Wilson KC, Bonham CA, Morgenthau AS, Patterson KC, et al.
491 Diagnosis and detection of sarcoidosis. an official american thoracic society clinical
492 practice guideline. *Am J Respir Crit Care Med* 2020;201:e26–51.
493 doi:10.1164/rccm.202002-0251ST.
- 494 [4] Judson MA. The diagnosis of sarcoidosis. *Curr Opin Pulm Med* 2019;25:484–96.
495 doi:10.1097/MCP.0000000000000596.
- 496 [5] Szeto HC, Coleman RK, Gholami P, Hoffman BB, Goldstein MK. Accuracy of
497 computerized outpatient diagnoses in a Veterans Affairs general medicine clinic. *Am J*
498 *Manag Care* 2002;8:37–43.
- 499 [6] Pendergrass SA, Crawford DC. Using electronic health records to generate phenotypes
500 for research. *Curr Protoc Hum Genet* 2019;100:e80. doi:10.1002/cphg.80.
- 501 [7] Horsky J, Drucker EA, Ramelson HZ. Accuracy and Completeness of Clinical Coding
502 Using ICD-10 for Ambulatory Visits. *AMIA Annu Symp Proc* 2017;2017:912–20.
- 503 [8] Ungprasert P, Matteson EL, Crowson CS. Accuracy of Diagnostic Coding for
504 Sarcoidosis in Electronic Databases: A Population-Based Study. *Lung* 2017;195:713–5.
505 doi:10.1007/s00408-017-0054-x.
- 506 [9] Velarde KE, Romesser JM, Johnson MR, Clegg DO, Efimova O, Oostema SJ, et al. An
507 initiative using informatics to facilitate clinical research planning and recruitment in the
508 VA health care system. *Contemp Clin Trials Commun* 2018;11:107–12.
509 doi:10.1016/j.conctc.2018.07.001.
- 510 [10] *Sampling Essentials: Practical Guidelines for Making Sampling Choices - SAGE*
511 *Research Methods* n.d. <https://dx.doi.org/10.4135/9781452272047> (accessed December
512 23, 2020).
- 513 [11] Moller DR, Koth LL, Maier LA, Morris A, Drake W, Rossman M, et al. Rationale and
514 Design of the Genomic Research in Alpha-1 Antitrypsin Deficiency and Sarcoidosis
515 (GRADS) Study. *Sarcoidosis Protocol. Annals of the American Thoracic Society*
516 2015;12:1561–71. doi:10.1513/AnnalsATS.201503-172OT.
- 517 [12] Baughman RP, Teirstein AS, Judson MA, Rossman MD, Yeager H, Bresnitz EA, et al.
518 Clinical characteristics of patients in a case control study of sarcoidosis. *Am J Respir*
519 *Crit Care Med* 2001;164:1885–9. doi:10.1164/ajrccm.164.10.2104046.
- 520 [13] National Center for Veteran Analysis and Statistics. Department of Veterans Affairs
521 Statistics at A Glance (Updated 12/31/2019). February 2020. n.d.
522 https://www.va.gov/vetdata/docs/Quickfacts/Stats_at_a_glance_4_6_20.PDF (accessed
523 January 1, 2021).
- 524 [14] Mirsaeidi M, Machado RF, Schraufnagel D, Sweiss NJ, Baughman RP. Racial
525 difference in sarcoidosis mortality in the United States. *Chest* 2015;147:438–49.
526 doi:10.1378/chest.14-1120.
- 527 [15] Cozier YC, Berman JS, Palmer JR, Boggs DA, Serlin DM, Rosenberg L. Sarcoidosis
528 in black women in the United States: data from the Black Women’s Health Study.
529 *Chest* 2011;139:144–50. doi:10.1378/chest.10-0413.
- 530 [16] Rybicki BA, Major M, Popovich J, Maliarik MJ, Iannuzzi MC. Racial differences in
531 sarcoidosis incidence: a 5-year study in a health maintenance organization. *Am J*
532 *Epidemiol* 1997;145:234–41. doi:10.1093/oxfordjournals.aje.a009096.
- 533 [17] Brito-Zerón P, Kostov B, Superville D, Baughman RP, Ramos-Casals M, Autoimmune
534 Big Data Study Group. Geoepidemiological big data approach to sarcoidosis:

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

- 535 geographical and ethnic determinants. *Clin Exp Rheumatol* 2019;37:1052–64.
- 536 [18] Arkema EV, Cozier YC. Sarcoidosis epidemiology: recent estimates of incidence,
537 prevalence and risk factors. *Curr Opin Pulm Med* 2020;26:527–34.
538 doi:10.1097/MCP.0000000000000715.
- 539 [19] Minority Veteran Statistics n.d.
540 https://www.va.gov/HEALTH/EQUITY/Race_Ethnicity.asp?utm_content&utm_medium=email&utm_name&utm_source=govdelivery&utm_term (accessed January 1,
541 2021).
542
- 543 [20] Organization WH. The ICD-10 Classification of Mental and Behavioural Disorders:
544 Diagnostic Criteria for Research. New Edition. Geneva: World Health Organization;
545 1993.
- 546 [21] Scadding JG. Prognosis of intrathoracic sarcoidosis in England. *BMJ* 1961;2:1165–72.
547 doi:10.1136/bmj.2.5261.1165.
- 548 [22] Su R, Li MM, Bhakta NR, Solberg OD, Darnell EPB, Ramstein J, et al. Longitudinal
549 analysis of sarcoidosis blood transcriptomic signatures and disease outcomes. *Eur*
550 *Respir J* 2014;44:985–93. doi:10.1183/09031936.00039714.
- 551 [23] Zhou T, Zhang W, Sweiss NJ, Chen ES, Moller DR, Knox KS, et al. Peripheral blood
552 gene expression as a novel genomic biomarker in complicated sarcoidosis. *PLoS One*
553 2012;7:e44818. doi:10.1371/journal.pone.0044818.
- 554 [24] Maier LA, Crouser ED, Martin WJ, Eu J. Executive summary of the NHLBI workshop
555 report: leveraging current scientific advancements to understand sarcoidosis variability
556 and improve outcomes. *Annals of the American Thoracic Society* 2017;14:S415–20.
557 doi:10.1513/AnnalsATS.201707-563OT.
- 558 [25] Sharma H, Mao C, Zhang Y, Vatani H, Yao L, Zhong Y, et al. Developing a portable
559 natural language processing based phenotyping system. *BMC Med Inform Decis Mak*
560 2019;19:78. doi:10.1186/s12911-019-0786-z.
- 561 [26] Richesson R. Electronic Health Records-Based Phenotyping n.d.
- 562 [27] Paul DW, Neely NB, Clement M, Riley I, Al-Hegelan M, Phelan M, et al.
563 Development and validation of an electronic medical record (EMR)-based computed
564 phenotype of HIV-1 infection. *J Am Med Inform Assoc* 2018;25:150–7.
565 doi:10.1093/jamia/ocx061.
566

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

567 **FIGURE LEGENDS**

568

569 **Figure 1. STROBE flow chart. Selection criteria for probable sarcoidosis cases.**

570 Abbreviations: ATS, American Thoracic Society; ICD, International Classification of

571 Diseases; PA, Palo Alto; SF, San Francisco; VA, Veterans Affairs

572

573 **Figure 2. Organ involvement assessment for probable sarcoidosis with and without**
574 **confirmed biopsy.** Abbreviations: AV, atrioventricular; CNS, central nervous system; CT,
575 computed tomography; CXR, chest radiograph; DLCO, diffusing capacity for carbon
576 monoxide; EKG, electrocardiogram; ENT, ears, nose and throat; MRI, magnetic resonance
577 imaging; PET, positron emission tomography; PFT, pulmonary function test.

578

579 **Figure 3. Distribution of characteristics among patients with probable sarcoidosis with**
580 **and without confirmed biopsy.** Abbreviations: PFT, pulmonary function test; ICD,

581 International Classification of Diseases.

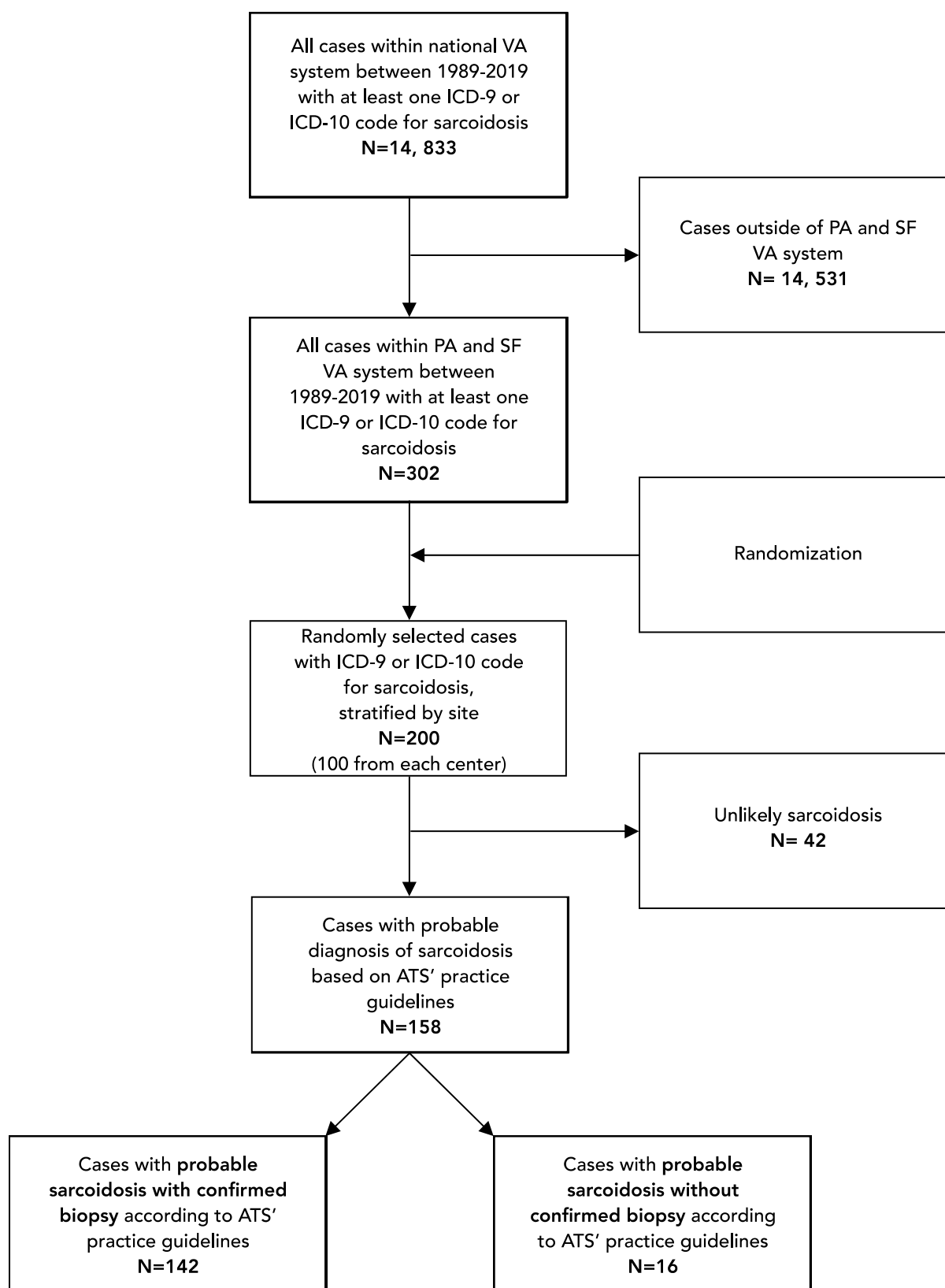
582

583 **Figure 4.** Comparison of specialty visits across cases with probable sarcoidosis (with and
584 without confirmed biopsy) and with healthcare providers classified as “Fully” or
585 “Substantially” compliant vs. “Partially” or “Non-” compliant using our compliance scoring
586 system.

587

588 **Acknowledgment:** The Author(s) confirmed no figures or tables included from another
589 publication.

590



591

592

593 **Figure 1.** STROBE flow chart. Selection criteria for probable sarcoidosis cases.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

594

Organ involvement	
Lung	<ul style="list-style-type: none"> • Positive lung biopsy, positive mediastinal/hilar lymph node biopsy • Chest X-ray/CT scan/PET scan demonstrating bilateral hilar lymphadenopathy, CT scan with perilymphatic nodules tracking the bronchovascular bundle, chest X-ray/CT scan/PET scan with diffuse infiltrates, chest X-ray with fibrosis • PFT with restriction and low DLCO or isolated reduced DLCO
Skin	<ul style="list-style-type: none"> • Positive skin biopsy • Lupus pernio and erythema nodosum ^a
Eye	<ul style="list-style-type: none"> • Positive conjunctival or scleral biopsy • Optic neuritis, scleritis, uveitis, or retinitis
Cardiac	<ul style="list-style-type: none"> • Positive heart/pericardium biopsy • 12-lead EKG showing Mobitz II or III-degree AV node block • AV node block or cardiomyopathy responsive to treatment • Cardiac MRI or PET-CT consistent with sarcoidosis
Liver	<ul style="list-style-type: none"> • Positive liver biopsy with a positive biopsy from another organ • Evidence of hepatomegaly • Abnormal liver enzymes
Multiorgan involvement	<ul style="list-style-type: none"> • More than two organs involved based on other criteria in this table
Neurosarcoidosis	<ul style="list-style-type: none"> • Positive brain/dura/peripheral nerve biopsy • Clinical syndrome or symptoms consistent with CNS sarcoidosis along with a positive MRI
ENT	<ul style="list-style-type: none"> • Positive biopsy from ear, nose, or throat • Direct laryngoscopy consistent with granulomatous disease

595

596

597

598

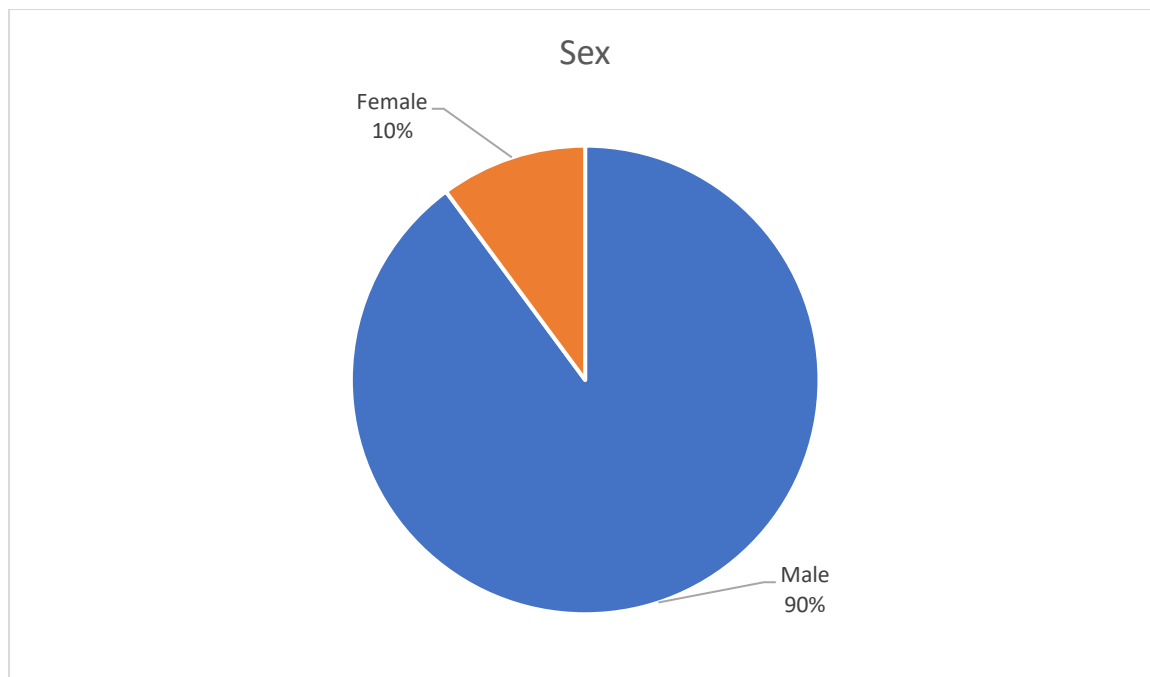
599

600

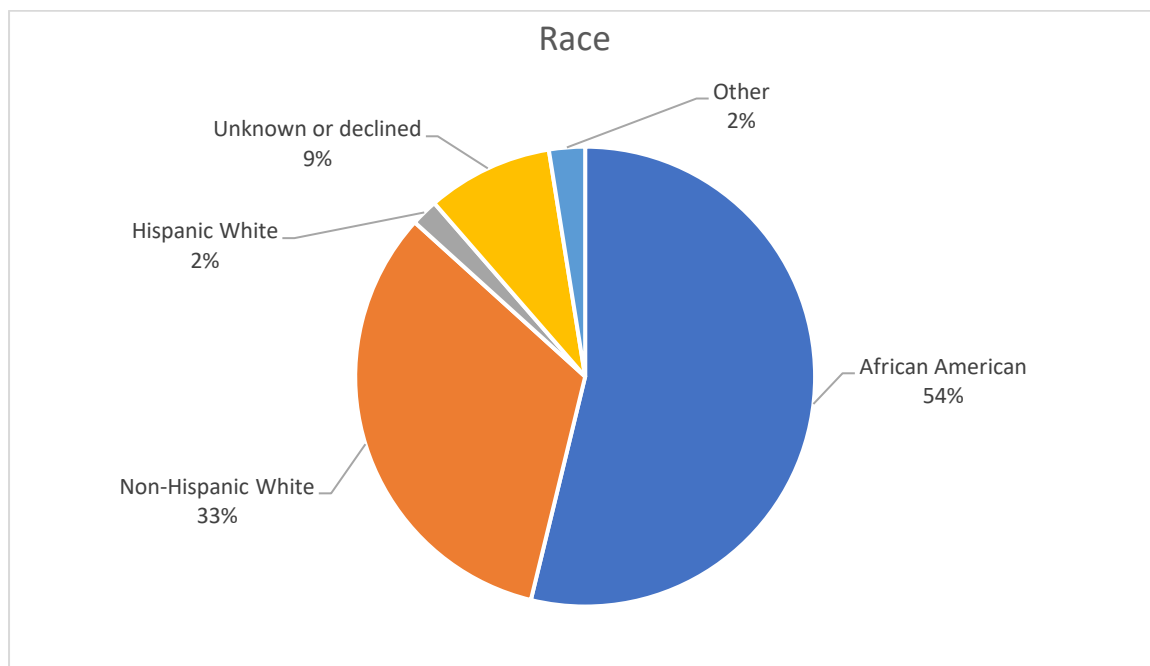
Figure 2. Organ involvement assessment for probable sarcoidosis with and without confirmed biopsy ^a.

^a No biopsy is needed for acute skin sarcoidosis.

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

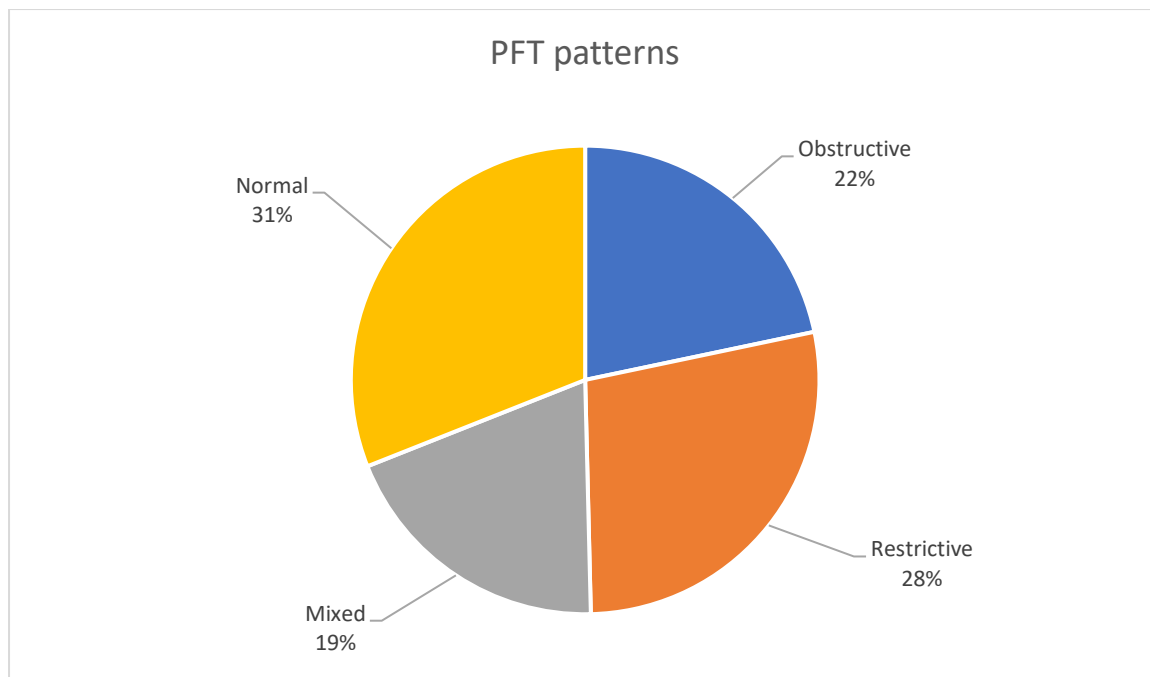


601
602

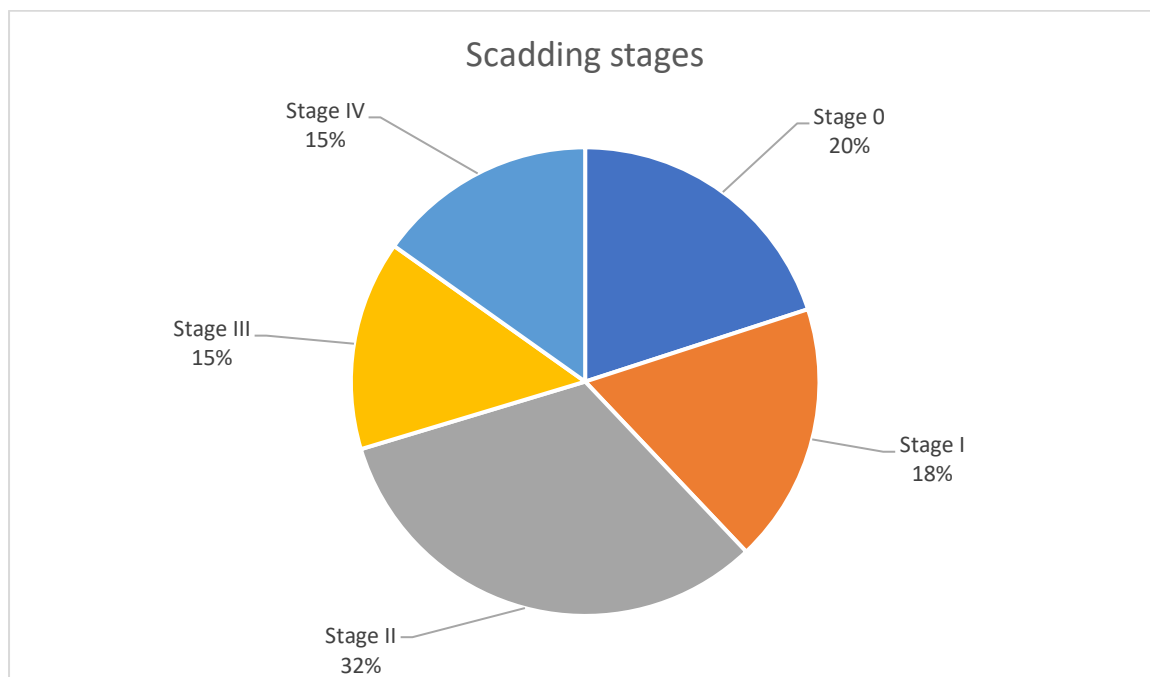


603

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

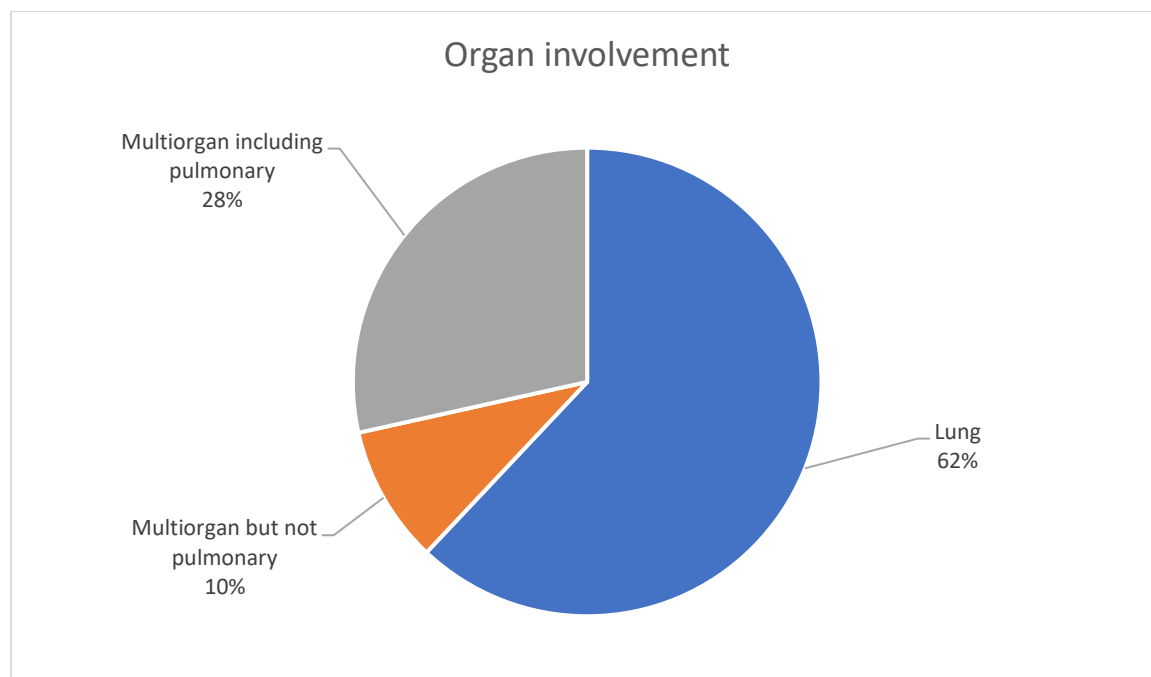


604
605

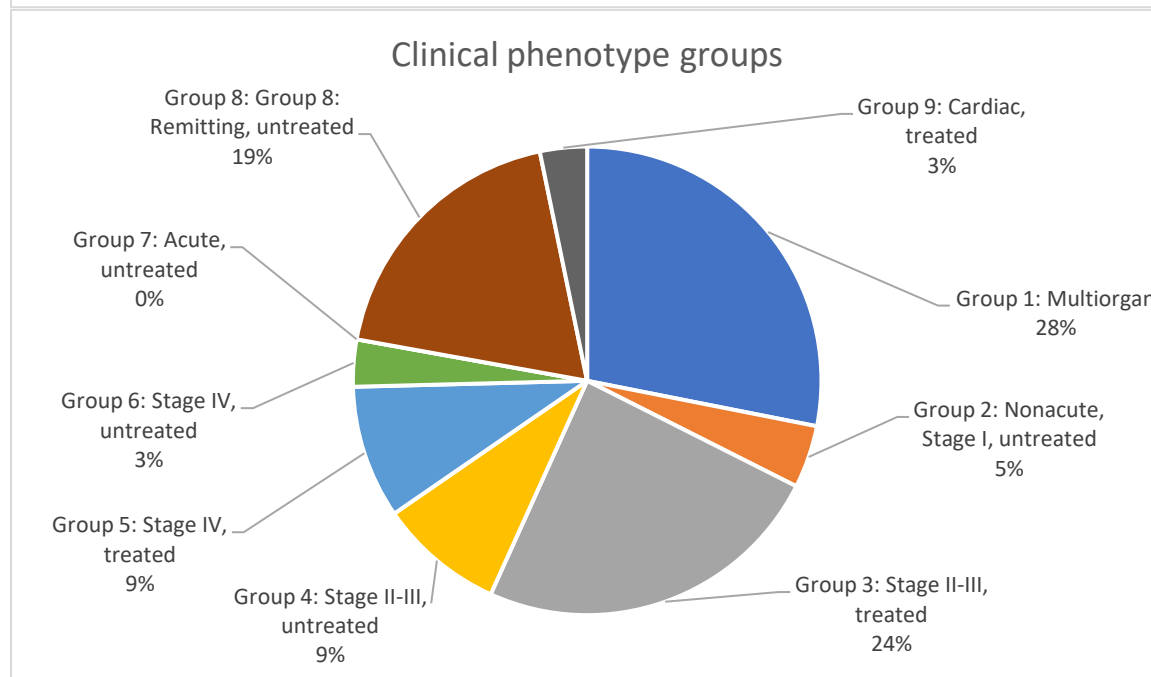


606

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

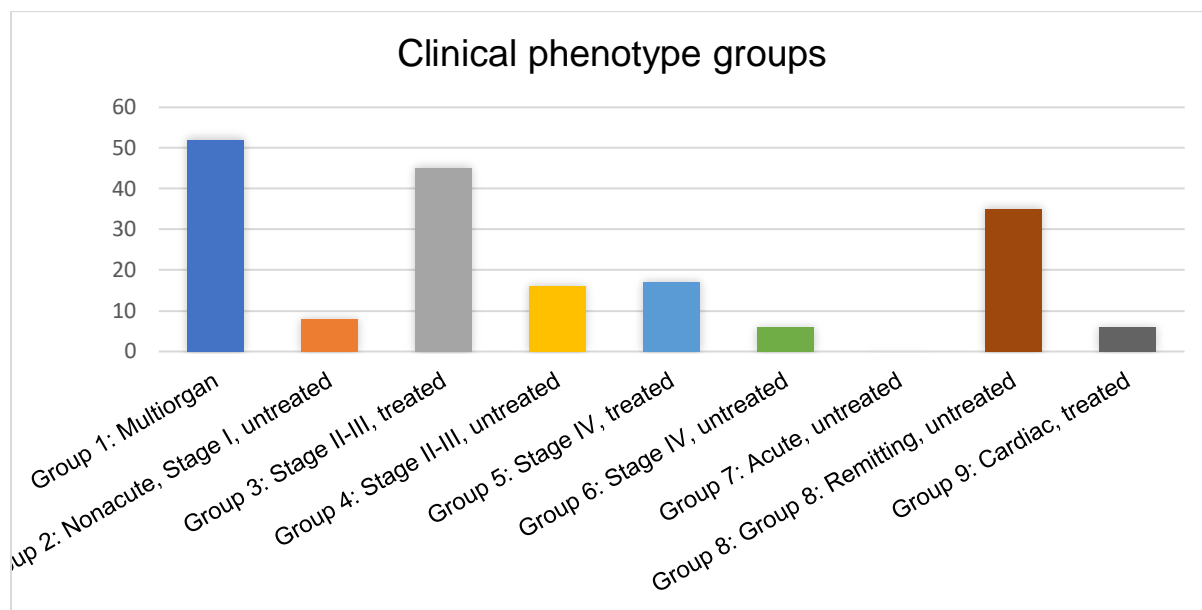


607

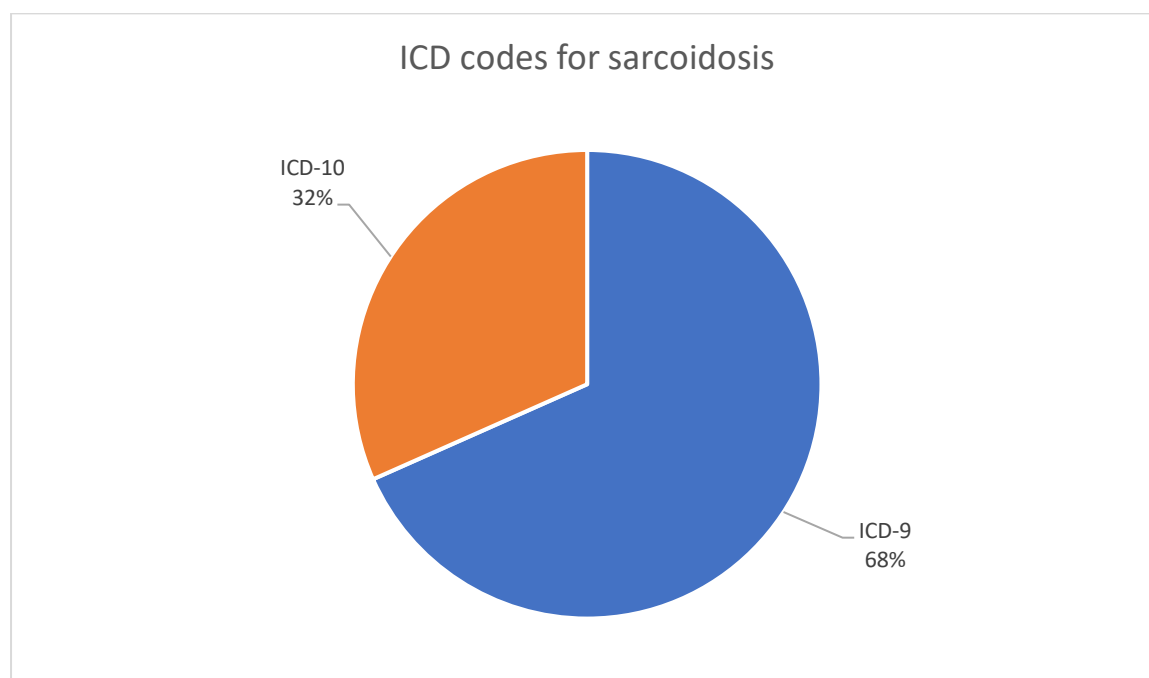


608

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)

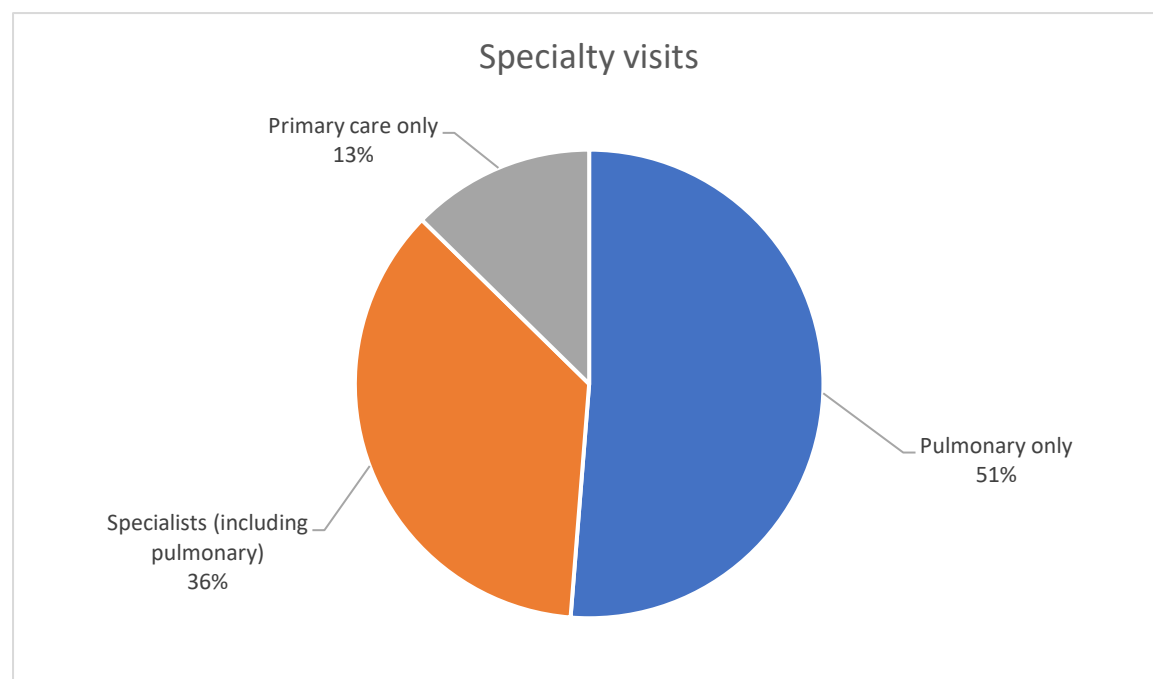


609
610



611
612
613
614
615

Mohamed I Seedahmed (<https://orcid.org/0000-0002-7446-7346>)



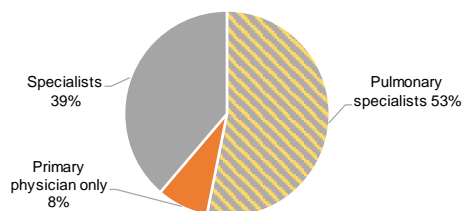
616

617 **Figure 3.** Distribution of characteristics among patients with probable sarcoidosis with and
618 without confirmed biopsy.

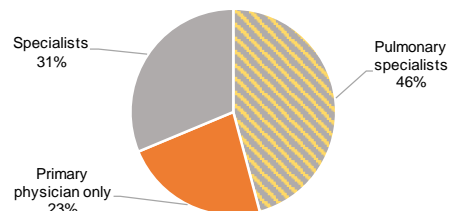
619 Abbreviations: PFT, pulmonary function test; ICD, International Classification of Diseases

620

Specialty visits among cases with **probable sarcoidosis (with and without confirmed biopsy)** and with providers' compliance score **"Fully"** or **"Substantially"** compliant



Specialty visits among cases with **probable sarcoidosis (with and without confirmed biopsy)** and with providers' compliance score **"Partially"** or **"Non-"** compliant



621

622

623 **Figure 4.** Comparison of specialty visits across cases with probable sarcoidosis (with and
624 without confirmed biopsy) and with healthcare providers classified as “Fully” or
625 “Substantially” compliant vs. “Partially” or “Non-” compliant using our compliance scoring
626 system.