BMJ Public Health

Estimating rare disease prevalence and costs in the USA: a cohort study approach using the Healthcare Cost Institute claims data

Christine M Cutillo,[1] Ainslie Tisdale,[1] Mahdi Baghbanzadeh,[2] Keith A Crandall ,[2] Reva L Stidd,[3] Manpreet S Khural,[3] Laurie J Hartman,[3] Jeff Greenberg,[3] Kevin B Zhang,[3] Ali Rahnavard [2]

To cite: Cutillo CM, Tisdale A, Baghbanzadeh M, *et al*. Estimating rare disease prevalence and costs in the USA: a cohort study approach using the Healthcare Cost Institute claims data. *BMJ Public Health* 2024;**2**:e000346. doi:10.1136/bmjph-2023-000346

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjph-2023-000346).

CMC, AT and MB contributed equally.

Received 26 June 2023
Accepted 17 January 2024

[1]National Center for Advancing Translational Sciences Office of Rare Diseases Research, Bethesda, Maryland, USA
[2]Department of Biostatistics and Bioinformatics, The George Washington University, Washington, District of Columbia, USA
[3]Customer Value Partners, Fairfax, Virginia, USA

Correspondence to
Dr Ali Rahnavard;
rahnavard@gwu.edu

ABSTRACT

Objective The study capitalised on national insurance claims data to gather information on patient characteristics and associated costs to better understand the diagnosis and treatment of rare diseases (RDs).

Materials and methods Data from the Healthcare Cost Institute (HCCI) data enclave were analysed using R statistical software and filtered by the International Classification of Diseases, 10th edition (ICD-10), current procedural terminology codes and the National Drug Code associated with 14 RDs and disease-modifying therapy options. Data were aggregated by prevalence, costs, patient characteristics and effects of treatment modification.

Results The prevalence and costs of RDs in the HCCI commercial claims database varied significantly across the USA and between urban and rural areas. Pharmacy costs increased when a new treatment was initiated, while non-pharmacy costs decreased.

Discussion Prevalence and cost estimations are highly variable due to the small number of patients with RDs, and the lack of a national healthcare database limits inferences for such patient populations. Accurate assessments require a diverse population, which can likely be achieved by analysing multiple databases. RDs face challenges in prevalence estimation due to a lack of specific disease coding and a small patient population, compounded by issues like data standardisation and privacy concerns. Addressing these through improved data management in healthcare systems, increased research and education will lead to better diagnosis, care management and quality of life for patients with RD.

Conclusion Data on patients with RD in the HCCI database were analysed for prevalence, costs, patient characteristics and treatment modification effects. Significant heterogeneity in each of these factors was found across RDs, geography and locality (eg, urban and rural). Building capabilities to use machine learning to accelerate the diagnosis of RDs would vastly improve with changes to healthcare data, such as standardising data input, linking databases, addressing privacy issues and assigning ICD-10 codes for all RDs, resulting in more robust data for RD analytics.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Rare diseases (RDs) affect a large number of people (30 million in the USA alone) and comprise vulnerable populations facing challenges in diagnosis and treatment.

WHAT THIS STUDY ADDS

⇒ The results from this study illustrate the significant heterogeneity in prevalence, cost and treatment effects among the RD population and suggest that strategies to improve data standardisation and physician or healthcare system (HCS) understanding of the impact of RDs will improve our collective ability to diagnose and treat patients with RD.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study's results are necessary for increasing our ability to accurately and expeditiously diagnose and treat patients with RD. RDs represent an urgent and extensive public health need currently unmet, with patients and HCS bearing the brunt of that burden. These study results, along with long-recognised considerations discussed in the RD community, illustrate the need and potential to incorporate tools—including such machine learning-based computational approaches—to leverage existing HCS data. Extension, expansion and validation of the processes performed for this study (in addition to mitigating the described limitations) indicate the ability to identify patients with RD sooner and more accurately and estimate the repercussions of undiagnosed RDs on HCS and patients, thereby decreasing the burden of RDs on patients, physicians and the entire HCS.

INTRODUCTION

Background and significance

The Orphan Drug Act defines a rare disease (RD) as a disease that impacts fewer than 200 000 people in the USA (https://rarediseases.info.nih.gov/about). However, with an estimated 7000–10 000 different RDs, 30 million—or 1 in every 10—Americans are collectively impacted by RDs.[1–5] Accordingly, when all RDs are considered together, they have staggering implications for a large swath of the USA and global population, on healthcare systems (HCS) and, most importantly, on patients with RD.[6]

Despite the number of Americans estimated to be impacted by RDs, many of these patients experience difficulty obtaining timely and accurate diagnoses. Reasons for this include unknown molecular mechanisms for diagnostics, a lack of US Food and Drug Administration (FDA)-approved treatments, difficulty in navigating patient data,[7] small and dispersed patient populations, diffused RD-specific expertise, overlapping symptoms and primary care physicians not well versed in all 10 000 RDs. As a result, many patients experience misdiagnosis and failed therapy interventions.[8 9] The path to diagnosis is often a prolonged journey that, according to the World Economic Forum,[10] can last an average of 7–8 years. Unfortunately, this lag in diagnosis may also result in missed opportunities to stop or slow RD progression. Patients may unknowingly forego disease-modifying therapy if available, or inappropriate care may be provided if misdiagnosed. Identifying individuals with RDs earlier could alleviate the long-term sequelae and financial burdens associated with RDs.[11 12] Patients with RD also grapple with limited treatment options, challenges finding a specialised physician or treatment centre, little or no research being conducted for their disease, high treatment costs and difficulty accessing medical, social or financial services or assistance.[13] Machine learning (ML), which encompasses a set of methodologies designed to gain insights and understanding from complex datasets,[14] offers an opportunity to better characterise RD and potentially lead to earlier diagnosis by identifying key features associated with RD. However, ML approaches require large volumes of data to be most effective in making predictions. Thus, it is critical for the successful applications of such approaches to access and use large national databases for RD research, as regional approaches will be limited in patient numbers for any given RD.

Objectives

The objective of this study is to comprehensively analyse and characterise the prevalence, patient characteristics and economic implications associated with RDs in the USA. The fragmented nature of the US HCS can lead to considerable variability in patient care and utilisation,[15] particularly concerning RDs. To address this, we have undertaken a detailed investigation using national commercial claims data sourced from the Healthcare Cost Institute (HCCI). This database encompasses a timeframe from 2012 to 2020 and includes data on an annual addition of 55 million individuals.

Building upon the groundwork laid by the Impact of Rare Diseases on Patients and Healthcare Systems (IDeaS) pilot study[5] conducted by the Division of Rare Disease Research Innovation within the National Institutes of Health National Centre for Advancing Translational Sciences (NCATS), our focus centres on 14 specific RDs. These conditions were the subject of the aforementioned pilot study, which encompassed diverse HCSs characterised by variations in geographic coverage, insurance representation, patient volume and duration of coverage. Our study aims to expand upon the insights garnered from the IDeaS pilot study by employing records exclusively from the HCCI database spanning the years 2016 through 2020. The primary objectives of this investigation are to estimate the prevalence of RDs across the USA, categorise the distribution of RDs within urban and rural communities, delineate the economic costs associated with RDs, classify patients based on their demographic and clinical characteristics and explore the impacts of treatment modifications. Through this multifaceted analysis, we seek to provide a comprehensive overview of the landscape of RDs in the USA, thereby contributing valuable insights to the understanding of these conditions and their implications for patients and HSCs.

METHODS

Population

In this study, we aim to build upon the pilot study conducted by Tisdale *et al*.[5] Our goal is to assess the prevalence, characteristic behaviours and costs associated with 14 RDs across the USA. To ensure consistency, we employed the same 14 RDs as in the pilot study, facilitating direct comparison of results.

Data collection

To achieve our objectives, we used data from the HCCI commercial claims database,[16] spanning the period from 2016 to 2020. This dataset enabled us to identify patients diagnosed with RDs and subsequently address specific questions concerning prevalence, characteristics and costs. To determine the prevalence, we employed a chronological approach, identifying patients with two instances of RD International Classification of Diseases, 10th edition (ICD-10) diagnosis codes,[17] identical codes or codes suggesting the same RD within a 3-month timeframe. For the purpose of stratification, we classified data based on urban–rural distinctions using zip codes from HCCI members and the Rural-Urban Commuting Area (RUCA) codes[18] provided by the US Department of Agriculture.

Our analysis also encompassed a detailed examination of costs related to RDs. We stratified costs by urban or rural categories and claim type, including inpatient, outpatient (facility charges), physician services and pharmacy expenses. Furthermore, we investigated the impact of disease-modifying therapies on patients' overall costs and claim frequencies. Specifically, we focused on cystic fibrosis (CF), studying the distribution and trends of claims and costs across different claim types before and after the initiation of treatment. For the analysis of treatment impact, we included only patients with a history of claims 90 days prior to and after treatment initiation.

Data analysis

The estimation of the prevalence of an RD involved determining the ratio of individuals diagnosed with the RD to the total patient count within the HCCI database for the designated timeframe. For comparisons among groups, a Kruskal-Wallis test along with pairwise Mann-Whitney tests were employed to probe for significant differences. We chose non-parametric tests such as the Mann-Whitney test due to their robustness to data distribution and outliers, unlike parametric tests like the t-test.[19 20]

For our cluster analysis, we aimed to produce the same number of clusters of patients for each disease. The primary objective of the analysis was to gain insights into the underlying structure of the data in an exploratory manner. As the dataset pertains to 14 RDs with potentially complex relationships between patients' clinical characteristics, a clustering approach was employed to identify any discernible patterns or similarities among the patients. Given the exploratory nature of the study and the absence of prior knowledge about the exact number of latent clusters, it was decided to use k-means clustering with five clusters as an initial step to partition the patients into distinct groups for further investigation. We clustered the patients for each of the 14 RDs separately based on claims per patient (CPP), the total number of CPPs and the average number of days between CPPs.

All the analyses described above were carried out using R[21] or R Studio within the HCCI data enclave. For a more comprehensive description of the methods used, please refer to online supplemental file 1. Additionally, the scripts employed for the various analyses are publicly accessible on our GitHub repository: https://github.com/omicsEye/rare_disease_claims.

Patient and public involvement

There were no patients involved in this study.

RESULTS

RD prevalence estimations

The HCCI database population (figure 1) consisted of a total of 117 908 879 members registered or active between 2016 and 2020, with 170 472 unique patients with one of the 14 RDs (0.145%). In the HCCI database, the RDs with the highest prevalence were eosinophilic esophagitis (EOE), sickle cell disease (SCD), CF, pheochromocytoma (Pheo) and muscular dystrophy (MD) (figure 2A).

This correlated with the top five most prevalent RDs in the medical literature, which were MD, EOE, SCD, hereditary haemorrhagic telangiectasia (HHT) and CF. The primary differences were the rank order and the symmetric difference of Pheo and HHT (figure 2B). The diseases with the lowest prevalence in the HCCI database were Batten disease (BD), mitochondrial neurogastrointestinal encephalopathy (MNGIE) and Takayasu's arteritis (TA), while the RDs with the lowest prevalence in the medical literature were MNGIE, Pheo and TA (online supplemental STable 1).

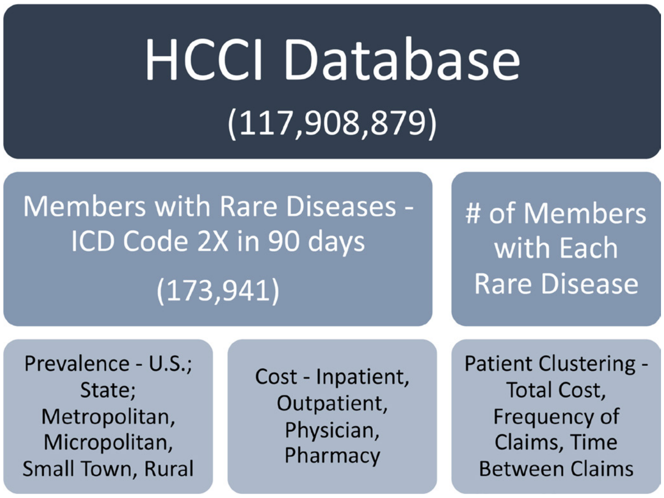
HCCI Database
(117,908,879)
Members with Rare Diseases - ICD Code 2X in 90 days
(173,941)
# of Members with Each Rare Disease
Prevalence - U.S.; State; Metropolitan, Micropolitan, Small Town, Rural
Cost - Inpatient, Outpatient, Physician, Pharmacy
Patient Clustering - Total Cost, Frequency of Claims, Time Between Claims

Figure 1 Schematic of data analysis in the HCCI database with patient numbers from 2016–2020.

Analysis of the prevalence of the 14 RDs at the state level and the RUCA level revealed prevalence was highly variable by state, yet there were no immediately discernible patterns (online supplemental STable 2). We determined statistically significant differences existed between the four region types for all diseases, warranting post hoc analyses using the Kruskal-Wallis test on the prevalence percentages of RUCA groups (online supplemental STable 3). The biggest geographical differences in prevalence ratios were (1) metropolitan to rural for SCD at 3.54:1, (2) small town to rural for TA at 3.5:1 and (3) metropolitan to rural for MNGIE at 2.25:1. Overall, the rural RUCA appeared to have a lower prevalence of RDs than other areas.

The implementation of pairwise Mann-Whitney tests for each RUCA combination within each RD revealed statistically significant differences for all pairs except for five: (1) CF small town and micropolitan, (2) TA small town and micropolitan, (3) BD small town and micropolitan, (4) MNGIE small town and micropolitan and (5) MNGIE small town and rural (online supplemental STable 4).

Estimations of total cost of RDs

The average CPP for each of the 14 RDs was higher than the control group, consisting of patients without RDs who had wellness visits (current procedural terminology (CPT) codes 90750, 90751, 90752 or 90754) (table 1). The costs varied widely across RDs, with CF having the highest average CPP, followed by urea cycle disorder (UCD), Lennox Gastaut syndrome (LGS) and MNGIE. EOE had a very similar CPP to the control population (table 1) and the highest prevalence in the HCCI data (figure 2A). The breakdown of total cost by claim type revealed that CF had the highest percentage of the total cost attributed to pharmacy claims at 59%, while pharmacy was the smallest

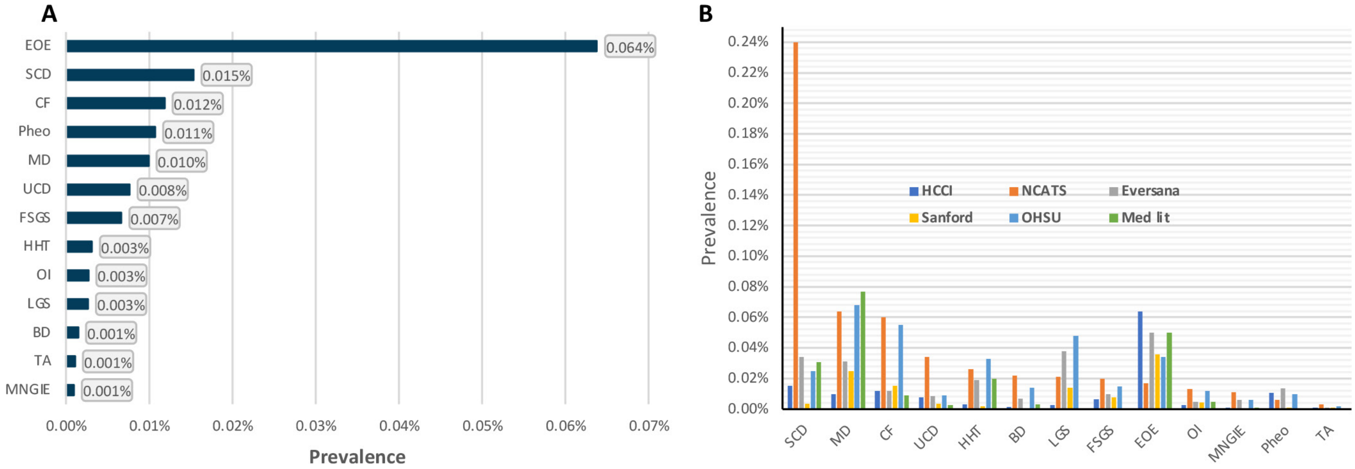

Figure 2 (A) Prevalence of RD patients across the U.S. based on HCCI commercial claims data 2016–2020. Eosinophilic esophagitis (0.064%) was the most prevalent, and mitochondrial neurogastrointestinal encephalopathy (0.001%) was the least prevalent in the HCCI population among the 14 RDs investigated in this study. (B) Estimated RD prevalence from different sources, including the National Center for Advancing Translational Sciences (NCATS), Oregon Health and Science University (OHSU), medical literature/public data sources (Med Lit), and Health Care Cost Institute (HCCI). HCCI was used as the source of analysis in this study, and the remaining values were from Tisdale et al.[5]

component of total costs for most other RDs (online supplemental SFigure 1). UCD had the highest inpatient cost component; focal and segmental glomerulosclerosis (FSGS) had the highest outpatient cost component, and MD had the highest physician cost component. All RDs, except EOE, had higher inpatient costs than the control group, and physician costs were slightly lower than those of the control group (online supplemental SFigure 1).

K-means clustering

For each disease, even with the cost variance among RDs shown earlier, there was always one distinct group, Cluster 5 (see online supplemental file 1). In half of the RDs, Cluster 5 made up the smallest portion of the patient population. In the other half, Cluster 5 was a slightly larger percentage of the patient population but never

Table 1 Total costs, number of patients and CPP for each RD (2016–2020)

RD	Total cost	Number of patients*	CPP
Cystic fibrosis	$3 451 710 078	13 965	$247 169
Eosinophilic esophagitis	$2 603 832 169	75 208	$34 622
Urea cycle disorder	$2 157 781 586	9005	$239 620
Sickle cell disease	$1 208 443 938	18 054	$66 935
Muscular dystrophy	$1 082 608 223	11 692	$92 594
Pheochromocytoma	$923 341 746	12 626	$73 130
Focal and segmental glomerulosclerosis	$824 118 245	7847	$105 023
Lennox Gastaut syndrome	$702 306 659	3114	$225 532
Charcot Marie tooth	$681 625 766	11 497	$59 287
Hereditary haemorrhagic telangiectasia	$223 485 801	3682	$60 697
Mitochondrial neurogastrointestinal encephalopathy	$203 847 961	1116	$182 659
Osteogenesis imperfecta	$181 155 862	3167	$57 201
Batten disease	$178 950 413	1704	$105 018
Takayasu's arteritis	$133 058 568	1264	$105 268
All RDs	$14 320 352 563	170 472	$84 004
Control	$23 483 336 297	695 463	$33 766

Total cost is the sum of all costs (inpatient, outpatient, physician and pharmacy) of patients diagnosed with that RD. All RDs show the aggregated values for unique individuals diagnosed with RD.
*Patients diagnosed with the RD.
CPP, cost per patient; RD, rare disease.

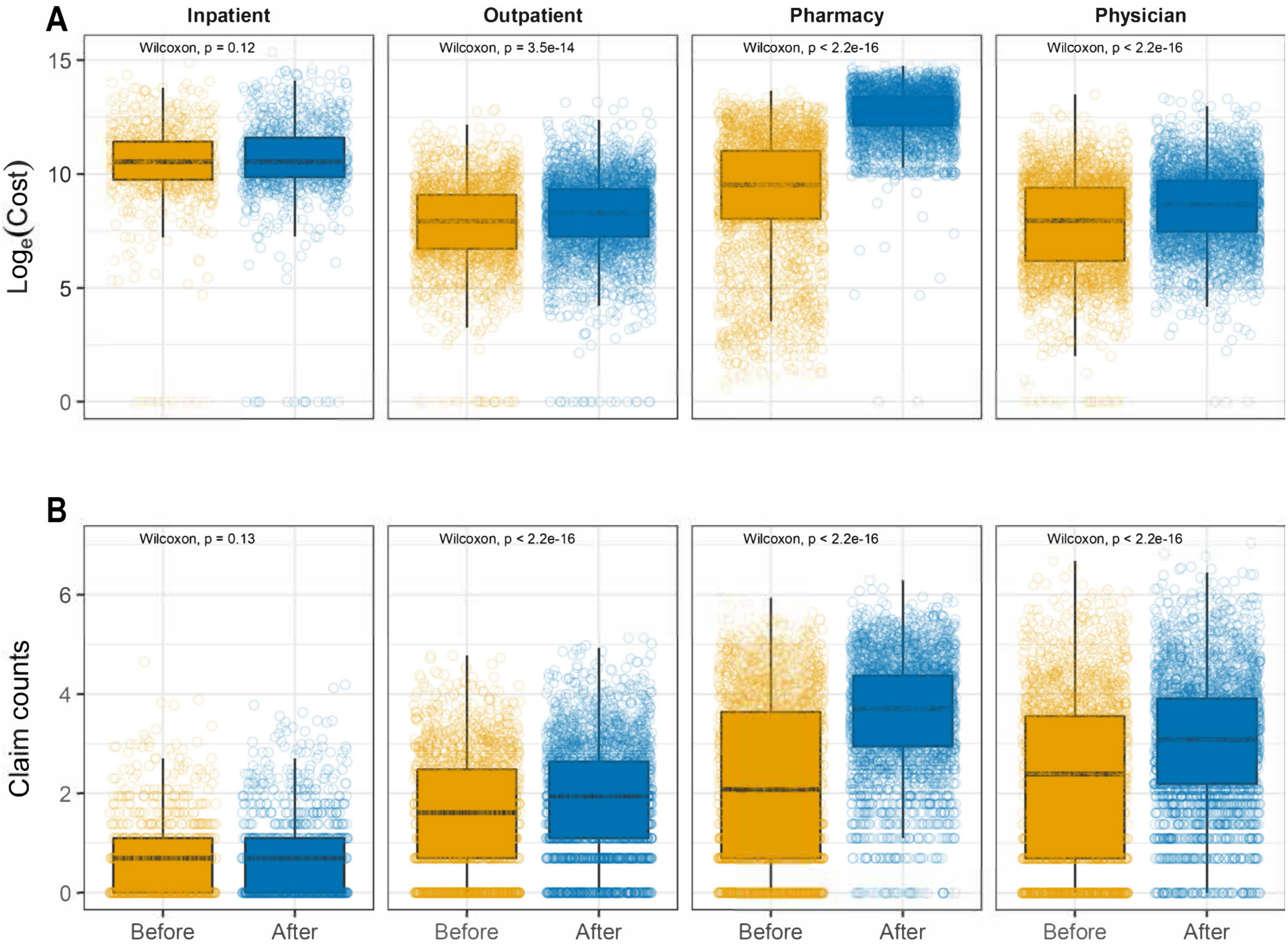

Figure 3 Distribution of (A) the natural log of total costs, and (B) claim counts of all 3372 cystic fibrosis patients with at least one claim record before and after starting Ivacaftor.

the largest (online supplemental STable 5). It contained the patients with the highest total costs, the most claims and the smallest amount of time between claims. We anticipated that some patients, perhaps those in need of critical care, would have a high utilisation of healthcare resources. Our findings validated this expectation and suggested that the remaining groups may have had some unique, discernible qualities as well.

Treatment modification effect

Cystic fibrosis

Ivacaftor is a disease-modifying therapy for patients with CF with particular mutations in the CF transmembrane conductance regulator (*CFTR*) gene (primarily the G551D mutation), which accounts for 4%–5% of all cases of CF.[22 23] Prior to initiating treatment with ivacaftor, the physician must confirm that the patient has one or more mutations in their *CFTR* gene, as well as adequate organ health and function to safely use this medication.

For the first CF subset cohort of 3372 patients, total claim counts and costs for these patients were higher after ivacaftor initialisation for all claim types except inpatient (figure 3A,B). The caveat here was that for CF, inpatient cost makes up 24% of total cost versus 17% of outpatient and physician combined (online supplemental SFigure 1).

To further examine this increase, a cost breakdown by yearly intervals before and after treatment commencement was performed, revealing similar non-pharmacy costs for most years except for the year before and the year after, during which there was a sizeable dip and spike in costs, respectively (online supplemental SFigure 2). These aberrations may be due to the clinical course of ivacaftor prescription. In the last year prior to the ivacaftor prescription, the patients' clinical care teams could have learnt enough regarding a patient's condition to perform more targeted tests than before, resulting in reduced non-pharmacy costs.

In the second subset of the CF cohort, we examined the 1375 patients with more extensive histories of claims before and after the initial medication prescription (figure 4A,B). Analysis of total, pharmacy only and non-pharmacy claims showed statistically significant differences before and after for both the number of claims per year and the cost per year (online supplemental STable 6). As compared with the previous cohort, the non-pharmacy claims decreased in both count and cost

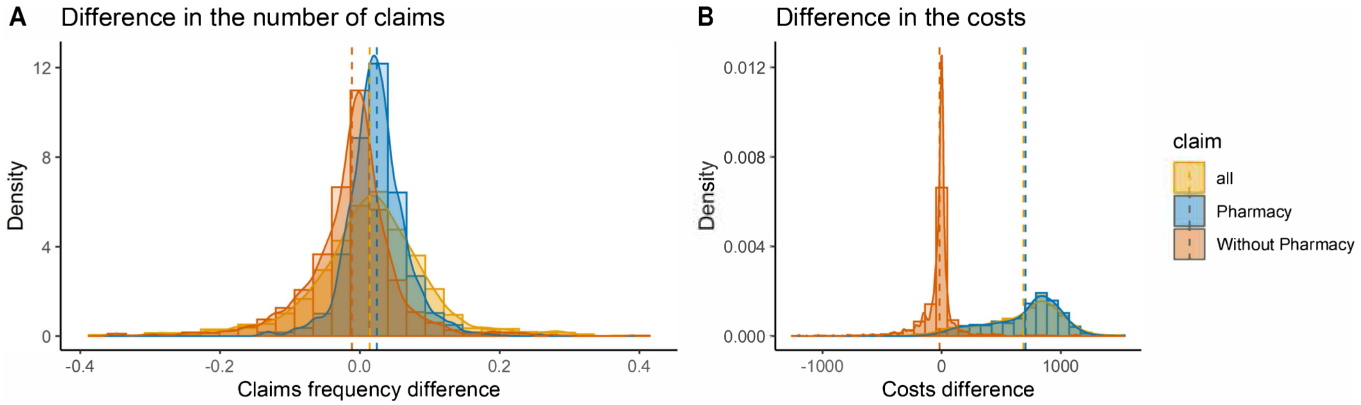

Figure 4 Distribution of the difference in (A) frequency of claims and (B) normalized costs by year between before and after using Ivacaftor. Only CF patients who have records of at least 90 days before and after taking this medication were considered in this analysis. The total number of patients is 1375.

by about four claims a year and \$6600 a year, respectively. These decreases were not enough; however, compared with the increases in pharmacy claims and costs, they were about \$256 700 more per year than before treatment. Additional analyses of other RDs showed similar trends with increased pharmacy costs.

DISCUSSION

This analysis builds on the previous pilot study by Tisdale *et al*[5] and expands upon it by using a different HCS database, HCCI and determining the prevalence and costs of RDs, patient clusters and the effect of treatment on costs and claim counts. These data allow researchers, physicians and policymakers to understand the overall impact of RDs on this employer-sponsored insurance (ESI) patient population. The information may provide actionable data to examine cost, improve management of RDs, initiate clinical trials and provide opportunities for machine learning for timely diagnosis of RDs.

Prevalence

In this study, we analysed 14 diverse RDs in the HCCI database for prevalence, costs and patient clusters. We used ICD-10 and CPT codes to identify members with the 14 selected RDs and determine the prevalence rates in this population. We demonstrated that reliable prevalence rates of RDs are difficult to determine as they are highly variable across databases. Several factors may contribute to this variability in prevalence estimates, including missing ICD codes and the characteristics of patient populations.

The data were stratified geographically by state. Prevalence rates in the various states showed no remarkable patterns, further demonstrating the dispersive nature of these diseases. Additional delineation based on zip code to assess rural versus urban populations revealed, as anticipated, that there is generally a lower prevalence in rural locales. This may be a result of potentially fewer healthcare resources, including facilities and specialty physicians, in rural areas. It could lead to patients relocating to more populated and resource-dense areas, prolong their diagnostic journey and make it harder to receive an accurate RD diagnosis. Additionally, the proportion of total rural workers who have ESI instead of other insurance plans is typically lower than the proportion of urban workers.[24] Patients with RD, in general, may be less likely to have ESI, skewing the prevalence estimate for rural areas. The prevalence of RUCA might vary due to several underlying factors. The dominantly significant differences found by the pairwise Mann-Whitney tests support the need to investigate further what may be leading to these variances. Exploring cost differences by RUCA may provide more insight into whether the prevalence rates truly represent the populations.

Costs

Costs for the 14 selected RDs varied widely, with EOE showing similarities to the control group costs. One possible explanation for the similar costs among patients with EOE may be due to a lack of treatment options or simpler disease management. Other RDs had significantly higher costs, with some higher claim-type proportions explained by specific disease needs, such as high pharmacy costs for CF due to new and costly drug therapies. Inpatient costs for patients with UCD may be due to frequent dialysis services. Further investigation is needed to determine whether these types of procedures are driving high inpatient costs and whether they can be safely moved to outpatient settings to reduce the costs and burden associated with overnight stays. Exploring specific high-cost procedures for each RD could also yield improvements to quality of life or life-saving or cost-saving opportunities.

Clustering

In the cluster analysis, five clusters were produced using total cost, number of claims and time between claims, which were mostly similar across all 14 RDs, indicating the existence of subpopulations that require further exploration using other patient or cohort characteristics. For instance, Group 5 for all RDs may indicate a specific geographical area and age combination, implying easier access to specialty hospitals, clinics and physicians in metropolitan areas. Further investigation is needed to determine which factors differentiate these clusters. Refinement of the clustering and expansion of the feature space could lead to the discovery of more specific subpopulation clusters composed of patient characteristic combinations not previously considered. Additionally, clustering should be performed on the control population to determine if the five consistent clusters found for all RDs apply to all patients in the HCCI data.

Treatment modification effects

When patients with CF began treatment with ivacaftor, as expected, pharmacy costs increased. We did not anticipate outpatient and physician costs to increase as well. However, managing medications for patients with CF, particularly patients with CF on ivacaftor, requires close monitoring. Patients generally undergo increased laboratory testing and physician visits, leading to increased non-pharmacy costs. Additionally, the prescription is filled 13 times per year as the drug is packaged in 28-day supplies, which leads to a higher frequency of pharmacy claims. Once stabilised, non-pharmacy costs and claim frequency decline. Without these drugs, patients are more prone to infections, likely resulting in poor quality of life and outcomes.[25]

Limitations

Missing ICD-10 codes is one source for the undercalculation of prevalence rates. Over half of all RDs do not have an ICD-10 code,[26] which is used for medical documentation and billing purposes. In some cases, researchers can use other criteria to identify patients with an RD, but this is not always possible.[5] Another source for variable prevalence data is the patient population in the healthcare database. RDs have a high burden of cost and can hinder daily life and the ability to maintain employment. Many patients living with RDs may transition their health insurance to Medicaid in full or with supplementary care. The HCCI Database covers ESI only, and therefore the members with RDs may be lower than seen in the general population. This was demonstrated by Tisdale *et al*[5] with the NCATS database that included Florida Medicaid patients, which was enriched with members diagnosed with RDs.

CONCLUSIONS

We successfully demonstrated the feasibility of searching HCS databases to gather valuable information on patients with RD and their characteristics, as well as RD costs. Future studies will continue to expand on these processes and analyses and will allow us to explore the use of machine learning tools to diagnose RDs more quickly. However, RD research faces challenges, and there are several limitations that we must overcome. Many RDs lack ICD or CPT codes. This lack of coding, imprecise coding and a limited number of patients impacted by each of these diverse ailments all contribute to the difficulty in identifying patients with RD in data enclaves, estimating the true prevalence of RDs in each population and measuring the impact they have on the patients and HCS. Additional considerations are standardisation of data elements across databases, solutions for ethical and privacy considerations and methods for connecting large databases or conducting analyses with multiple databases to achieve a more diverse and representative population cohort.[27] Improving HCS data, increasing research, educating physicians and expanding drug development will ultimately lead to faster diagnoses, better management of care, and an improved quality of life for those affected by RDs.

Acknowledgements We thank Dr Andrea Hahn, Children's National Hospital, for her help in interpreting ICD and CPT codes for RDs.

Contributors AT, CMC, KAC, JG, RLS and AR conceived or designed the work. JG, RLS and KAC acquired data access. MB and AR led data analyses, and AT, CMC, RLS, KAC, MB, AR, MSK, KBZ and LJH interpreted the data for the work. All authors drafted the manuscript and revised it critically for important intellectual content, reviewed and approved the manuscript for submission, agreed to be accountable for all aspects of the work and ensure that questions of accuracy and integrity were investigated. The opinions expressed in this article are the authors' (AT and CMC) own and do not reflect the views of the National Institutes of Health. AR is the guarantor of the study.

Funding This work was supported by the National Institutes of Health/National Centre for Advancing Translational Sciences grant number 75N95021Q00316.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study on human subjects was approved by the George Washington University Committee on Human Research, Institutional Review Board (IRB, GW IRB NCR213979), FWA00005945, and waiver of consent approved by the IRB using de-identified claims data.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. All data relevant to the study are included in the article or uploaded as supplementary information. The data used in this article are available through a third party in the Healthcare Cost Institute (HCCI) data enclave, and the code used for data analysis is available on our GitHub repository: https://github.com/omicsEye/rare_disease_claims.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
Keith A Crandall http://orcid.org/0000-0002-0836-3389
Ali Rahnavard http://orcid.org/0000-0002-9710-0248

REFERENCES

1 Institute of Medicine, Board on Health Sciences Policy, Committee on Accelerating Rare Diseases Research and Orphan Product Development. *Rare diseases and orphan products: accelerating research and development*. National Academies Press, 2011.

2 Groft SC. Rare diseases: identifying needs. *American Pharmacy* 1990;30:33–40.

3 About - genetic and rare diseases information center. Available: https://rarediseases.info.nih.gov/diseases/pages/31/faqs-about-rare-diseases [Accessed 1 Aug 2022].

4 Nguengang Wakap S, Lambert DM, Olry A, *et al*. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet* 2020;28:165–73.

5 Tisdale A, Cutillo CM, Nathan R, *et al*. The IDeaS initiative: pilot study to assess the impact of rare diseases on patients and healthcare systems. *Orphanet J Rare Dis* 2021;16:429.

6 Whicher D, Philbin S, Aronson N. An overview of the impact of rare disease characteristics on research methodology. *Orphanet J Rare Dis* 2018;13:14.

7 Decherchi S, Pedrini E, Mordenti M, *et al*. Opportunities and challenges for machine learning in rare diseases. *Front Med (Lausanne)* 2021;8:747612.

8 Pogue RE, Cavalcanti DP, Shanker S, *et al*. Rare genetic diseases: update on diagnosis, treatment and online resources. *Drug Discov Today* 2018;23:187–95.

9 Haendel M, Vasilevsky N, Unni D, *et al*. How many rare diseases are there? *Nat Rev Drug Discov* 2020;19:77–8.

10 Nothaft W, Goldsmith C, Le Cam Y. It takes far too long for a rare disease to be diagnosed. Here's how that can change. World Economic Forum. 2020. Available: https://www.weforum.org/agenda/2020/02/it-takes-far-too-long-for-a-rare-disease-to-be-diagnosed-heres-how-that-can-change/ (accessed 1 Aug 2022

11 Biffi A, Montini E, Lorioli L, *et al*. Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. *Science* 2013;341:1233158.

12 Finkel RS, Mercuri E, Darras BT, *et al*. Nusinersen versus Sham control in infantile-onset spinal muscular atrophy. *N Engl J Med* 2017;377:1723–32.

13 National Organization for Rare Disorders. Rare disease day: frequently asked questions. Available: https://rarediseases.org/wp-content/uploads/2019/01/RDD-FAQ-2019.pdf [Accessed 18 Jul 2022].

14 James G, Witten D, Hastie T, *et al*. *An introduction to statistical learning*. New York, NY: Springer US, 2013.

15 Agha L, Frandsen B, Rebitzer JB. Fragmented division of labor and healthcare costs: evidence from moves across regions. *Journal of Public Economics* 2019;169:144–59.

16 Blewett LA, Call KT, Turner J, *et al*. Data resources for conducting health services and policy research. *Annu Rev Public Health* 2018;39:437–52.

17 Hirsch JA, Nicola G, McGinty G, *et al*. ICD-10: history and context. *AJNR Am J Neuroradiol* 2016;37:596–9.

18 Cromartie J. Rural-urban commuting area codes. Available: https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/ [Accessed 14 Sep 2022].

19 Zimmerman DW, Zumbo BD. Effect of outliers on the relative power of parametric and nonparametric statistical tests. *Percept Mot Skills* 1990;71:339–49.

20 MacFarland TW, Yates JM. Mann–Whitney U test. In: MacFarland TW, Yates JM, eds. *Introduction to nonparametric statistics for the biological sciences using R*. Cham: Springer International Publishing, 2016: 103–32.

21 Core Team R. *R: A language and environment for statistical computing Version 3.6*. Vienna, Austria, Available: /ra-language-and-environment-forstatistical-computing

22 Jones AM, Helm JM. Emerging treatments in cystic fibrosis. *Drugs* 2009;69:1903–10.

23 McPhail GL, Clancy JP. Ivacaftor: The first therapy acting on the primary cause of cystic fibrosis. *Drugs Today* 2013;49:253.

24 Newkirk V, Damico A. The affordable care act and insurance coverage in rural areas. 2014 Available: https://www.kff.org/uninsured/issue-brief/the-affordable-care-act-and-insurance-coverage-in-rural-areas/

25 Kirwan L, Fletcher G, Harrington M, *et al*. Longitudinal trends in real-world outcomes after initiation of ivacaftor. A cohort study from the cystic fibrosis registry of Ireland. *Ann Am Thorac Soc* 2019;16:209–16.

26 Strashny A, Alford J, Rappole C, *et al*. The National hospital care survey is a unique source of data on rare diseases. *Value in Health* 2022;25:1814–7.

27 Thompson R, Johnston L, Taruscio D, *et al*. RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. *J Gen Intern Med* 2014;29 Suppl 3:S780–7.