Data quality and patient characteristics in European ANCA-associated vasculitis registries: data retrieval by federated querying


Handling editor Josef Smolen

ABSTRACT

Objectives This study aims to describe the data structure and harmonisation process, explore data quality and define characteristics, treatment, and outcomes of patients across six federated antineutrophil cytoplasmic antibody-associated vasculitis (AAV) registries.

Methods Through creation of the vasculitis-specific Findable, Accessible, Interoperable, Reusable, VASCulitis ontology, we harmonised the registries and enabled semantic interoperability. We assessed data quality across the domains of uniqueness, consistency, completeness and correctness. Aggregated data were retrieved using the semantic query language SPARQL Protocol and Resource Description Framework Query Language (SPARQL) and outcome rates were assessed through random effects meta-analysis.

Results A total of 5282 cases of AAV were identified. Uniqueness and data-type consistency were 100% across all assessed variables. Completeness and correctness varied from 49%–100% to 60%–100%, respectively. There were 2754 (52.1%) cases classified as granulomatosis with polyangiitis (GPA), 1580 (29.9%) as microscopic polyangiitis and 937 (17.7%) as eosinophilic GPA. The pattern of organ involvement included: lung in 3281 (65.1%), ear-nose-throat in 2860 (56.7%) and kidney in 2534 (50.2%). Intravenous cyclophosphamide was used as remission induction therapy in 982 (50.7%), rituximab in 505 (17.7%) and pulsed intravenous glucocorticoid use in 24.8 (95% CI 19.7 to 31.1) per 1000 patient-years, respectively.

Conclusions In the largest reported AAV cohort-study, we federated patient registries using semantic web technologies and highlighted concerns about data quality. The comparison of patient characteristics, treatment and outcomes was hampered by heterogeneous recruitment settings.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Semantic web technologies enable simultaneous privacy preserving querying of multiple distributed data sources.

WHAT THIS STUDY ADDS

⇒ This is the first study to demonstrate successful implementation of federated data integration in a rheumatic disease.
⇒ This is the first study to address issues in completeness and correctness of European vasculitis registry data.
⇒ Comparison of patient characteristics, treatments and outcomes of antineutrophil cytoplasmic antibody-associated vasculitis (AAV) across European registries is presented in largest AAV cohort ever reported.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This first attempt to federate AAV patient registries shows the potential for international benchmarking of patient care using distributed registries with variable registry design.
⇒ This study highlights the need for structured assessment of data quality to allow for scientifically robust research and reliable decision-making.
⇒ Data source federation is not limited to clinical registries, the infrastructure may be readily expanded to include genomic and postgenomic research data and the methodology is readily scalable to other rheumatic diseases.

INTRODUCTION

Antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis (AAV) is a group of systemic autoimmune disorders characterised by inflammation and destruction of predominantly small blood
the Polish Vasculitis Registry (POLVAS),18 Ireland’s Rare Kidney Disease (RKD) registry19 and Sweden’s Skåne Vasculitis Cohort.1

vessels. AAVs encompass three disease subtypes, sharing clinical and pathological features and association with ANCA; granulomatosis with polyangiitis (GPA), microscopic polyangiitis (MPA) and eosinophilic GPA (EGPA). All three types of AAV are rare, with a combined annual incidence rate of about 30 per million adults.3 A sufficiently large number of observations are required for reliable statistical inference of patient data, but low incidence and consequently small sample sizes of observational cohorts present a major barrier to clinical research in AAV.

A growing number of clinical trials in vasculitis, and the subsequent need to collect long-term data on treatment safety and efficacy in routine care, has led to development of several vasculitis registries across Europe.2 3 These registries play an increasing role in research, patient care and healthcare planning. In the FAIRVASC (Findable, Accessible, Interoperable, Reusable, VASculitis) project, AAV data from six well-established European vasculitis registries are federated.3 The project has been developed in the context of the European Joint Programme on Rare Diseases (EJPRD), a Europe-wide initiative to overcome challenges related to data fragmentation in rare disease research.3 Using a semantic-web approach, the challenge of heterogeneous data structures and semantics have been addressed by the creation of a vasculitis-specific FAIRVASC ontology, as well as aligning the project with the FAIR principles of scientific data management, making data findable, accessible, interoperable and reusable.6–8

To allow valid and reliable decision-making based on scientific data, it is important to ensure data are of high quality. The literature proposes a diverse range of dimensions and methodologies to describe and measure data quality.9–11 However, the quality of European vasculitis registries or vasculitis research data has rarely been formally assessed.

Our aims in this observational retrospective cohort study are to (1) describe the data structure, federation and harmonisation process, (2) explore data quality and (3) give an overview on the baseline characteristics, treatment and outcomes of AAV patients across six European registries.

**METHODS**

**Registry federation**

This descriptive study is part of a data reuse project aiming to federate multiple existing vasculitis registries. Federation stands in contrast to traditional centralisation where multiple local independent datasets are combined in a central data pool for analysis. Instead, data are held within the firewall of its home institution to be queried and analysed remotely, addressing data privacy, access and security concerns inherent in establishing a central data pool. Through federation local registries retain control over data access rights in compliance with their ethical approvals and/or informed consent of registry participants. In FAIRVASC, data federation is enabled through a web interface,12 following previous approaches for medical data linkage.13 14 Participant registries keep autonomous control over how registry data are represented internally, while exposing an endpoint to allow federated retrieval of non-subject level data through the website (figure 1).

**The registries**

The participating registries are the Czech Registry of AAV,15 the French Vasculitis Study Group registry (FVSG),16 the Joint Vasculitis Registry in German-speaking Countries (GeVas),17 the Polish Vasculitis Registry (POLVAS),18 Ireland’s Rare Kidney Disease (RKD) registry19 and Sweden’s Skåne Vasculitis Cohort.1

These registries were selected through the registries working group of the European Vasculitis Society (EUVAS) as a combination of highly mature and recently established registries and reflecting a balance of rheumatology and nephrology-led initiatives. A key criterion was the availability of outcome data (eg, death and end-stage kidney disease (ESKD)), thereby excluding some important registries such as UKIVAS. The characteristics of the registries, including location and period of recruitment, tools for data collection and follow-up, are summarised in table 1.

We included all patients with a diagnosis of AAV, including GPA, MPA, EGPA and unspecified AAV, regardless of classification. Two (Czech and RKD) of the registries excluded patients with a classified diagnosis of AAV s that were positive for anti-glomerular basement membrane antibodies. In four registries (FVSG, Skåne and POLVAS, RKD), unspecified AAVs were excluded or not recruited.

**Data harmonisation**

The data structure and variables collected across the registries are different. To enable semantic interoperability, we harmonised the definitions of data elements and their corresponding values and converted registry data into a structure that adheres to a harmonised schema. This was achieved by the creation of the vasculitis-specific FAIRVASC ontology, that is, a representation, formal naming and definition of the categories, properties and relations between the concepts, data and entities within the registries. In the conversion to the harmonised schema, we uplifted heterogeneous registry data to a machine-readable format using the declarative mapping language ‘RDB to RDF Mapping Language’ and enriched data across the registries by adding semantic meaning to data using terms from standardised ontologies, such as the Orphanet Rare Disease Ontology.20 The use of ontologies is a key part of the FAIR principles of scientific data management and stewardship, an initiative putting emphasis on machine readability and reusability of data.7 21 The development of the FAIRVASC ontology and semantic integration

Figure 1 FAIRVASC project workflow. Registry data are quality controlled and harmonised to a common schema. Harmonised data are uploaded to a local server with an exposed endpoint. A researcher dispatches a predefined query from a web interface to the local servers. Statistical calculations are performed locally, and non-subject level results sent back. Data are combined and presented to the web-interface user. FAIRVASC, Findable, Accessible, Interoperable, Reusable, VASculitis.
of vasculitis data have previously been described in detail by McGlinn et al. This harmonisation process results in transformation of unstructured relational data to a knowledge graph data format.

Data quality

We assessed data quality in the registries across four core domains: uniqueness, consistency, completeness and correctness. These domains were selected in collaboration with the European Institute for Innovation through Health Data (i~HD), based on a pool of nine candidate dimensions in the i~HD data quality assessment framework. In assessing uniqueness (absence of duplication of data), we investigated duplicate patient identifiers and potential cases of readmission of the same patient under another unique identifier within each registry. Consistency was assessed by examining nine key variables in AAV research (patient gender, date of birth, immunoassay ANCA-type, Birmingham Vasculitis Activity Score (BVAS) at diagnosis, serum creatinine at diagnosis, C reactive protein at diagnosis, induction treatment, date of death and date of ESKD) for the appropriate data type according to the registries’ data dictionaries. Logic tests were applied to variables containing dates (eg, ‘Was date of death greater than date of diagnosis’). Plausibility tests of numeric values were performed for two variables (eg, ‘Was serum creatinine at diagnosis within a biologically plausible range of 0–5000 µmol/L’). For the completeness domain, the amount of missing data was assessed across the same variables. Correctness of data entry of the variables described above was assessed in each registry against an electronic health record for at least ten patients per registry. Based on these four domains, we developed a data quality worksheet (online supplemental file 1). Using this worksheet, data quality was assessed at each registry site by local investigators. The results of the data quality assessment are presented as variable specific percentages, stratified by registry.

Table 1  Meta-data summary of the FAIRVASC registries

<table>
<thead>
<tr>
<th></th>
<th>Czech*</th>
<th>FVSG</th>
<th>GeVas</th>
<th>POLVAS</th>
<th>RKD</th>
<th>Skåne</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current no of AAV cases</td>
<td>335</td>
<td>2804</td>
<td>169</td>
<td>936</td>
<td>668</td>
<td>374</td>
</tr>
<tr>
<td>Types of AAV*</td>
<td>EGPA, MPA, GPA, Unspecified AAV</td>
<td>EGPA, MPA, GPA, Unspecified AAV</td>
<td>EGPA, MPA, GPA, Unspecified AAV</td>
<td>EGPA, MPA, GPA, Unspecified AAV</td>
<td>EGPA, MPA, GPA</td>
<td>EGPA, MPA, GPA</td>
</tr>
<tr>
<td>Recruitment area</td>
<td>National</td>
<td>National</td>
<td>National</td>
<td>National</td>
<td>National</td>
<td>Regional</td>
</tr>
<tr>
<td>Name of recruitment area</td>
<td>Czech Republic</td>
<td>France</td>
<td>Germany</td>
<td>Poland</td>
<td>Ireland</td>
<td>Skåne, Sweden</td>
</tr>
<tr>
<td>Dominant medical specialties</td>
<td>Nephrology</td>
<td>All</td>
<td>All</td>
<td>All</td>
<td>Nephrology</td>
<td>All</td>
</tr>
<tr>
<td>Study design</td>
<td>Mixed retrospective/ prospective</td>
<td>Prospective cohort study</td>
<td>Mixed retrospective/ prospective</td>
<td>Retrospective cohort study</td>
<td>Mixed retrospective/ prospective</td>
<td>Mixed retrospective/ prospective</td>
</tr>
<tr>
<td>Follow-up</td>
<td>Longitudinal</td>
<td>Longitudinal</td>
<td>Longitudinal</td>
<td>Longitudinal</td>
<td>Longitudinal</td>
<td>Longitudinal</td>
</tr>
<tr>
<td>Data source</td>
<td>Encounter based</td>
<td>Encounter based</td>
<td>Encounter based</td>
<td>Chart review</td>
<td>Encounter based</td>
<td>Chart review</td>
</tr>
</tbody>
</table>

*In this study, only single-centre data starting from 2013 is presented. EMA, CHCC 2012, 2012 Revised International Chapel Hill Consensus Conference Nomenclature of Vasculitides; ACR 1990 criteria. AAV, anti-neutrophil cytoplasmic antibody-associated vasculitis; ACR, American College of Rheumatology; EGPA, Eosinophilic granulomatosis with polyangiitis; EMA, European Medicines Agency; FAIRVASC, Findable, Accessible, Interoperable, Reusable, Vascularitis; FVSG, French Vasculitis Study Group; GeVas, Joint Vasculitis Registry in German-speaking Countries; GPA, Granulomatosis with polyangiitis; MPA, Microscopic polyangiitis; POLVAS, Polish Vasculitis Registry; RKD, Rare Kidney Disease.

Patient characteristics retrieval

We analysed patients recruited to the FAIRVASC registries with a diagnosis of AAV up to 31 January 2023, with descriptive analysis of demographics, type of diagnosis, organ involvement, serology, serum creatinine level at diagnosis, treatments, follow-up and outcome. Continuous variables are summarised with a mean and SD. Categorical variables are summarised as frequencies. We estimated the incidence rate of ESKD and all-cause mortality rate combined across the registries per 1000 person-years with a 95% CI using random effects meta-analyses. The definition of ESKD differed slightly between the registries (online supplemental file 2). The aggregated patient characteristics were retrieved from the knowledge graphs of the registries through the semantic query language SPARQL (online supplemental file 1), and random effects meta-analysis was performed using the R package metafor. As the federated queries did not allow for the flow of subject level data from one site to another or to a centralised server, the investigation of summary statistics requiring the global rank of data to be known (eg, quantiles) could not be assessed.

Patient and public involvement

The sharing of data in this study concerns aggregated patient data and is thus not governed by general data protection regulations. As health data are sensitive and rare disease data difficult to fully anonymise, we, in this study, chose to federate aggregated registry data, as opposed to develop and transfer data to a centralised data pool. This allows for sustainability through real-time updates of data presented, rigorous data access control by the data owner and scalability to include additional mature and emerging vasculitis registries. Patients and patient advocacy organisations have been involved in the design and conduct of the FAIRVASC project from inception. This ensures consideration of patient priorities and perspectives regarding the ethics, privacy and confidentiality of the federated research approach,
the design of this study and future communication and dissemination of this work.

RESULTS

Patient characteristics and treatment

A total of 5282 patients (2568 (48.6%) women) were included across the 6 registries. Of these, 2734 (52.1%) were classified as GPA, 1580 (29.9%) MPA and 937 (17.7%) as EGPA according to different sets of international classification criteria or clinical diagnosis. The mean age at diagnosis was 56.0±16.7 years. There were 1840 (51.2%) PR3-ANCA positive, 1506 (41.8%) MPO-ANCA positive and 219 (9.0%) immunoassay ANCA negative patients. Pulmonary involvement was reported in 3281 (65.1%), ear-nose-throat involvement in 2860 (56.7%) and kidney involvement in 2534 (50.2%). Cardiovascular and abdominal involvement were present in 822 (16.3%) and 658 (13.4%), respectively. The mean creatinine level at diagnosis was 198±266 μmol/L across the registries (table 2).

Data quality

No duplicate patient identifiers were identified in the registries, although, in one registry, there were 2.2% potential duplicate entries (ie, whereby the same patient was entered under more than one unique identifier). Consistency of data type was 100% across all assessed variables present in the registries. Consistency for logic tests of dates was between 95.6% and 100% and between 98.7% and 100% for plausibility tests for numeric variables. Completeness for gender and date of birth ranged from 95.1% to 100% across all registries. Not all variables were present in all registries due to differences in registry design. Completeness for laboratory tests at diagnosis ranged from 49.5% to 99.2%. A BVAS assessment was available at diagnosis for 49.5%–100% of patients. In the registries which record induction treatment, at least one treatment type was recorded in 96.9% to 100% of the cases. In patients who died or reached ESKD by the end of follow-up, a date of death or date of ESKD was available in 75%–100%. Data correctness was 60.0%–100% across all variables and registries (figure 2A–C and online supplemental file 1).

Treatment

Some information on induction treatment was collected in all registries, but the granularity of information differed. Cyclophosphamide could be assessed in five of six registries, and oral and intravenous use could be separated in four registries. Information about rituximab was present in five of six registries. The use of pulsed intravenous glucocorticoids and plasma exchange could be assessed in all registries. Oral and intravenous cyclophosphamide was used in 357 (18.4%) and in 1716 (60.0%), respectively, and rituximab was used in 505 (17.7%) patients. Plasma exchange was used in 421 (13.4%) and pulsed intravenous glucocorticoids in 1735 (55.2%). Maintenance treatment was recorded in 4 of 6 registries, with azathioprine used in 911 (50.1%) and rituximab in 282 (15.5%) (table 2).

Outcome analysis

The mean follow-up time was 6.2±5.8 years with a total of 30 548 person-years of follow-up. During follow-up, there were 767 deaths occurring with a known date of death, yielding a pooled all-cause mortality rate of 28.8 (95% CI 19.7 to 42.2) per 1000 patient-years (figure 3). The highest mortality rates were seen in the Skåne and Czech registries, 62.8 (95% CI 54.4 to 72.4) and 40.5 (95% CI 30.0 to 54.5), respectively. The pooled incidence-rate of ESKD was 24.8 (95% CI 19.7 to 31.1) per 1000 patient years (figure 4). There was considerable heterogeneity among the registries (I²=96% and I²=75%, respectively). The I² statistic describes the percentage of variation across studies that is due to heterogeneity rather than chance.

DISCUSSION

In this observational retrospective cohort study, the largest yet reported, we describe the data structure and harmonisation process, explore data quality and give an overview on baseline characteristics and outcomes of AAV patients across six European patient registries within the FAIRVASC project.

Despite collaborative efforts in recent years, research data in vasculitis have remained fragmented, siloed, and is rarely standardised from the perspective of data interoperability. To allow federated analytics across multiple registries, a process of data harmonisation across data sources is needed. The FAIRVASC ontology provides a framework for harmonisation that is scalable to existing or emerging AAV registries and cohorts. The use of semantic web technologies in rare disease research data integration has precedents in the medical literature. Importantly, the RD-Connect project and EJPRD Virtual Platform are high-level data integration projects aiming to achieve interoperability over a wide range of rare diseases and data resources. Our development of a disease-specific ontology provides a framework suitable for clinical research. The cohort presented in this study is already of unprecedented size in AAV and will soon be extended by the addition of further AAV registries using the framework provided.

High-quality data are essential to allow reliable decision-making. Despite this, there are no formalised frameworks for assessment of data quality specific for rare disease research and no agreed thresholds recognised for what constitutes sufficient data quality. We assessed data quality based on the i−HD data quality assessment framework. Arguably, the critical threshold for data quality varies depending on what data will be used for. This study demonstrated high data quality regarding uniqueness and consistency across participating registries. However, laboratory values and disease activity assessments at diagnosis lacked in completeness. A correctness assessment, comparing a subset of registry data with that of health records, showed some non-matching information in all but one registry.

This highlights evident data quality issues in present day vasculitis registries. Lack of completeness and some missing data are an inevitable part of any real-life dataset and can be addressed statistically. Additionally, the completeness assessment does not investigate why data were missing and may, in some cases, reflect registry design rather than completeness. For example, a creatinine value at baseline may appear to be missing but would reasonably not be collected in a patient on acute dialysis. Clinical data collection is inherently subjective which may be reflected in the completeness assessment. BVAS has previously shown high interobserver reliability by Suppiah et al, but the study was limited by small numbers. Selecting an appropriate laboratory test value in relation to a date of diagnosis may also be subject to interpretation. Potential steps to improve data quality in vasculitis registries include further enhancement of data validation systems, automated input from electronic health records and targeted data collection to improve completeness and correctness of key variables.

This cohort of 5282 AAV patients is the largest cohort of AAV ever reported. GPA was the most common diagnosis across the registries, while MPA and EGPA followed. The varying classification criteria used for registry inclusion and the wide temporal
**Table 2** Patient characteristics summary stratified by registry

<table>
<thead>
<tr>
<th></th>
<th>Czech</th>
<th>FVSG</th>
<th>GeVas</th>
<th>POLVAS</th>
<th>RKD</th>
<th>Skåne</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demography</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age at diagnosis, mean (SD)*</td>
<td>63.1</td>
<td>61.1</td>
<td>63.6</td>
<td>63.6</td>
<td>63.9</td>
<td>63.6</td>
<td>63.6</td>
</tr>
<tr>
<td>Men, n (%)</td>
<td>162</td>
<td>135</td>
<td>90</td>
<td>62</td>
<td>63</td>
<td>40</td>
<td>74</td>
</tr>
<tr>
<td>Women, n (%)</td>
<td>164</td>
<td>155</td>
<td>119</td>
<td>119</td>
<td>116</td>
<td>40</td>
<td>125</td>
</tr>
<tr>
<td><strong>Diagnosis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPA, n (%)</td>
<td>143</td>
<td>139</td>
<td>85</td>
<td>64</td>
<td>72</td>
<td>56</td>
<td>83</td>
</tr>
<tr>
<td>MPA, n (%)</td>
<td>76</td>
<td>64</td>
<td>46</td>
<td>43</td>
<td>38</td>
<td>27</td>
<td>42</td>
</tr>
<tr>
<td>EGA, n (%)</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Unspecified AAV, n (%)</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>ANCAs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PR3-ANCA positive, n (%)</td>
<td>43</td>
<td>40</td>
<td>23</td>
<td>23</td>
<td>21</td>
<td>15</td>
<td>23</td>
</tr>
<tr>
<td>MPO-ANCA positive, n (%)</td>
<td>30</td>
<td>26</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>ELISA negative, n (%)</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td><strong>S-creatinine at diagnosis, mean (SD)</strong>*</td>
<td>74</td>
<td>76</td>
<td>78</td>
<td>78</td>
<td>79</td>
<td>80</td>
<td>79</td>
</tr>
<tr>
<td><strong>Organ pattern involvement</strong>††</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General, n (%)</td>
<td>208</td>
<td>209</td>
<td>155</td>
<td>82</td>
<td>258</td>
<td>277</td>
<td>384</td>
</tr>
<tr>
<td>Mucocutaneous/eeye, n (%)</td>
<td>37</td>
<td>37</td>
<td>67</td>
<td>43</td>
<td>62</td>
<td>66</td>
<td>66</td>
</tr>
<tr>
<td>Ear-nose-throat, n (%)</td>
<td>112</td>
<td>112</td>
<td>94</td>
<td>62</td>
<td>112</td>
<td>112</td>
<td>112</td>
</tr>
<tr>
<td>Lung, n (%)</td>
<td>160</td>
<td>112</td>
<td>90</td>
<td>62</td>
<td>112</td>
<td>112</td>
<td>112</td>
</tr>
<tr>
<td>Cardiovascular, n (%)</td>
<td>7</td>
<td>17</td>
<td>12</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Abdominal, n (%)</td>
<td>10</td>
<td>11</td>
<td>11</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Kidney, n (%)</td>
<td>310</td>
<td>108</td>
<td>57</td>
<td>62</td>
<td>57</td>
<td>62</td>
<td>57</td>
</tr>
<tr>
<td>Nervous, n (%)</td>
<td>51</td>
<td>63</td>
<td>64</td>
<td>96</td>
<td>41</td>
<td>51</td>
<td>51</td>
</tr>
<tr>
<td>Induction treatment§§§</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oral cyclophosphamide, n (%)</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IV cyclophosphamide, n (%)</td>
<td>389</td>
<td>87</td>
<td>734</td>
<td>340</td>
<td>166</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Rituximab, n (%)</td>
<td>112</td>
<td>75</td>
<td>112</td>
<td>177</td>
<td>29</td>
<td>505</td>
<td>505</td>
</tr>
<tr>
<td>Plasma exchange, n (%)</td>
<td>88</td>
<td>44</td>
<td>81</td>
<td>137</td>
<td>62</td>
<td>421</td>
<td>421</td>
</tr>
<tr>
<td>Intravenous glucocorticoids, n (%)</td>
<td>258</td>
<td>122</td>
<td>699</td>
<td>367</td>
<td>41</td>
<td>1375</td>
<td>1375</td>
</tr>
<tr>
<td>Maintenance treatment††</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Azathioprine, n (%)</td>
<td>26</td>
<td>26</td>
<td>32</td>
<td>206</td>
<td>91</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>Methotrexate, n (%)</td>
<td>44</td>
<td>19</td>
<td>163</td>
<td>56</td>
<td>182</td>
<td>182</td>
<td>182</td>
</tr>
<tr>
<td>Mycophenolate mofetil, n (%)</td>
<td>0</td>
<td>0</td>
<td>122</td>
<td>51</td>
<td>345</td>
<td>345</td>
<td>345</td>
</tr>
<tr>
<td><strong>Outcome</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Death, n (%)</td>
<td>56</td>
<td>350</td>
<td>113</td>
<td>127</td>
<td>187</td>
<td>836</td>
<td>836</td>
</tr>
<tr>
<td>ESKD, n %††††</td>
<td>49</td>
<td>298</td>
<td>145</td>
<td>127</td>
<td>55</td>
<td>679</td>
<td>679</td>
</tr>
<tr>
<td>Follow-up time in years, mean (SD)§§§</td>
<td>3.4</td>
<td>6.4</td>
<td>6.1</td>
<td>7.3</td>
<td>7.9</td>
<td>6.2</td>
<td>6.2</td>
</tr>
</tbody>
</table>

If not commented complete data. All percentages presented as valid percentages.

*Available for n: 334 (Czech), 2529 (FVSG), 169 (GeVas), 901 (POLVAS), 666 (RKD), 374 (Skåne), 4973 (Total).
†Not included in registry collection.
‡Excluded in registry harmonisation.
§Available for n: 333 (Czech), 1283 (FVSG), 169 (GeVas), 774 (POLVAS), 668 (RKD), 374 (Skåne), 3601 (Total).
¶Not available information in registry design.
**Available for n: 253 (Czech), 2017 (FVSG), 138 (GeVas), 509 (RKD), 371 (Skåne), 3288 (Total).
††Cases on acute or chronic dialysis actively excluded
‡‡Available for n: 332 (Czech), 2574 (FVSG), 168 (GeVas), 929 (POLVAS), 668 (RKD), 372 (Skåne), 5043 (Total).
§§Available for n: 283 (Czech), 761 (FVSG), 149 (GeVas), 922 (POLVAS), 661 (RKD), 365 (Skåne), 3141 (Total).
†††Information not available due to registry design.
‡‡‡Cannot differ administration route of cyclophosphamide in registry design. Presented is all cyclophosphamide.
††††Available for n: 69 (GeVas), 857 (POLVAS), 563 (RKD), 331 (Skåne), 1820 (Total).
†††End-stage kidney disease definitions: Czech (dialysis for >90 days; sustained CKD 5 (estimated glomerular filtration rate (eGFR)<15 mL/min/1.73 m²) for >90 days; and/or kidney transplantation), FSVG (dialysis for more than 30 days or death within 30 days of start of dialysis), GeVas (renal replacement therapy; sustained dialysis or CKD 5 in two succeeding visits), Skåne/POLVAS (sustained dialysis), RKD (dialysis for >90 days; sustained CKD 5 (eGFR<15 mL/min/1.73 m²) for >90 days; and/or kidney transplantation).
§§§Available for n: 203.
AAV, antineutrophil cytoplasmic antibody associated vasculitis; EGPA, eosiinophilic granulomatosis with polyangiitis; ESKD, end stage kidney disease; FSVG, French Vasculitis Study Group; GeVas, Joint Vasculitis Registry in German-speaking Countries; GPA, granulomatosis with polyangiitis; MPA, microscopic polyangiitis; POLVAS, Polish Vasculitis Registry; RKD, rare kidney disease.
spread in the period of recruitment needs to be considered when interpreting these results. The low frequency of MPA in the FVSG registry is expected with registry inclusion starting in 1983 from a multidisciplinary perspective. MPA was recognised as a distinct disease subtype in the 1994 Chapel Hill Consensus Criteria.30 The high frequency of MPA in the Czech and RKD registries may be attributed to the nephrology-centred patient inclusion of these registries. The mean age at diagnosis in the Skåne registry was significantly higher than in the other registries. This can likely be explained by the Skåne registry’s population coverage design. This may indicate that, as in clinical trials, the oldest patient groups are underrepresented in patient registries.31 PR3 positivity was the most common ANCA pattern overall, largely following the frequency of diagnosis subtype, with MPO positivity being more common in registries recruiting primarily from a nephrology setting. We saw a large variance in organ pattern distribution of active disease across the registries, likely reflecting the different recruitment settings. These differences in disease phenotype need to be considered when assessing regional observational studies.

Intravenous cyclophosphamide was the most administered remission induction treatment, followed by oral cyclophosphamide and rituximab. The use of remission maintenance treatment could be assessed in four registries with azathioprine being...
For the potential future implementation of federated learning to enable joint modelling on data from all included registries. Seeking to address data governance and privacy concerns, federated learning approaches are, therefore, being developed to fit statistical models and algorithms (eg, the extended analysis of patient outcome, including adjusted proportional hazards models) to distributed datasets without exchanging the data itself. This is key in projects where data privacy concerns restrict data access and prohibits transfer of data to a centralised pool. A successful implementation of federated learning holds the potential for precision medicine on a large-scale while respecting privacy concerns, overcoming the limitations of approaches that require a single pool of centralised data.

The harmonisation schema developed within this project is readily extendable to existing and emerging AAV registries. This is simplified through the development of the EUVAS ‘Model registry’, providing a framework for the prospective collection of a defined set of data items. Working in conjunction with the FAIRVASC harmonisation schema, this is building the foundation for further data integration and interoperability of AAV registries. Data source federation is not limited to clinical registries and the infrastructure may be expanded to include genomic and postgenomic research data. The methodology is also scalable to other rheumatic diseases where research suffers from small sample sizes and governance hurdles of international data sharing.

In this observational retrospective cohort study, we present data from six well-established European AAV registries represented in knowledge graphs, supporting scalable federated benchmarking of care. A data quality assessment indicated high
levels of data quality regarding uniqueness and consistency of data, but there are data quality concerns regarding completeness and correctness for some of the key variables in AAV research. Our analysis of patient characteristics and outcomes of AAV across Europe is difficult to interpret due to differences in registry patient inclusion setting and period of recruitment. However, there is variability in patient care not adequately explained by these differences. This first attempt to federate AAV patient registries demonstrates the potential for benchmarking of patient care across Europe and the development of precision medicine at large-scale while respecting patient privacy concerns using a federated approach.

Author affiliations
1. Clinical Sciences, Rheumatology, Lund University, Lund, Sweden
2. School of Infectious and Immunity, University of Glasgow, Glasgow, UK
3. Department of Mathematical Science, University of Durham, Durham, UK
4. Telemedicine Technologies, Liège, Belgium
5. ADAPT SFI Centre, School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
6. Department of Nephrology, General University Hospital, Prague, Czech Republic
7. First Faculty of Medicine, Charles University, Prague, Czech Republic
8. National Referral Center for Rare Systemic Autoimmune Diseases, Hospital Cochin, Paris, France
9. Université Paris Cité, Paris, France
10. Department of Rheumatology and Clinical Immunology, Universitat zu Lubeck, Lubeck, Germany
11. Division of Molecular Biology and Clinical Genetics, Jagiellonian University Medical College, Krakow, Poland
12. 2nd Department of Internal Medicine, Jagiellonian University Medical College, Krakow, Poland
13. French Vasculitis Study Group, Paris, France
14. Trinity Kidney Centre, Trinity Translational Medicine Institute, Trinity College Dublin, Dublin, Ireland
15. Department of Clinical Sciences, Lund University, Lund, Sweden
16. Department of Endocrinology, Nephrology and Rheumatology, Skåne University Hospital, Lund, Sweden
17. General University Hospital in Prague, Prague, Czech Republic
18. Nephrology and Dialysis Unit, Meyer Children’s Hospital IRCCS, Firenze, Italy
19. Clinical Trials Unit, Medical Center, University of Freiburg Faculty of Medicine, Freiburg, Germany
20. School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
21. Department of Medicine, University of Cambridge, Cambridge, UK

Twitter Louis Aslett @louisaslett and Mark A Little @thkc1

Acknowledgements The authors would like to thank the patient organisations Vasculitis International and Vasculitis Ireland Awareness, and especially Peter Verhoeven and Julie Power for ensuring the consideration of patient priorities and perspectives throughout the project. Thanks to Nathan Lea, Maria Christofidou and Dipak Kalra at the European Institute for Innovation through Health Data for guidance and support in data governance issues. Thanks also to Pintal Ltd and Grace Davy, and Elena Gelain at Meyer Children Hospital Florence for administrative assistance and Professor Cathal Walsh at the School of Medicine, Trinity College Dublin for statistical guidance.

Collaborators The FAIRVASC consortium: Adrian Tassoni (Clinical Trials Unit, Medical Center—University of Freiburg); Aleatea Tais (Clinical Trials Unit, Medical Center—University of Freiburg); Cathal Walsh (School of Medicine, Trinity College Dublin); Cecil Armstrong (Vasculitis Ireland Awareness); Dipak Kalra (The European Institute for Innovation through Health Data); Elena Gelain (Nephrology and Dialysis Unit, Meyer Children’s Hospital IRCCS); Fabian Schubach (Interdisciplinary Medical Institute for Innovation through Health Data); Gabriele Ihorst (Clinical Trials Unit, Medical Center—University of Freiburg); Hansnette Aerts (The European Institute for Innovation through Health Data); Iris Sengers (Vasculitis International); John Mills (Vasculitis UK); Julie Power (Vasculitis Ireland Awareness); Katarzyna Wawrzyczka-Adamczyk (II Chair of Internal Medicine, Faculty of Medicine, Jagiellonian University Medical College); Kris McGlinn (School of Computer Science and Statistics, Trinity College Dublin); Margaret Dunne (Technological University of Dublin); Maria Christofidou (The European Institute for Innovation through Health Data); Nathan Lea (The European Institute for Innovation through Health Data); Peter Verhoeven (Vasculitis Stichting); Raissa de Boer (Rheumatology, Department of Clinical Sciences, Lund University); Sabrina Arnold (Department of Rheumatology and Clinical Immunology, University of Lübeck).

Contributors KG developed the SPARQL queries to retrieve the aggregated data and developed the figures and tables and vouch for the analyses of data. MR and MT directed and designed data quality collection. FD, DW, HK, SL, JS, RS and BY developed the FAIRVASC ontology and technical infrastructure and performed local data quality assessments. LH and DO’S oversaw the development of FAIRVASC technical infrastructure. NB, ZH, PL, JM, XP, MS, BT, AV and KW contributed to the interpretation of scientific data. LA and AW contributed to the statistical analyses. UIM and MAL directed the data quality assessment equally as last authors and take full responsibility for the finished work, the conduct of the study and controlled the decision to publish. All authors contributed intellectual content during the draft and revision of the work and approved the final version to be published.

Funding This project has received funding from the European Union’s Horizon 2020 2020 research and innovation programme under the EJP RD COFUND-EIP No 825575, Crafoordska stiftelsen: 20220623 (KG and A/JM). Vetenskapsrådet: 2019-00263 (KG and A/JM), Science Foundation Ireland 13/RC/2106_P2 and 11/Y/B2093 (MAL). European Union’s Horizon 2020 research and innovation programme under the EJP RD: EJP-RD/FAIRVASC/03/2020 (K/NCB/00058) (SL, JM and KWI). CSL Vifor; IIS-NA-2021-0497 (NB and MR). Fundings from the French Vasculitis Study Group who implemented and coordinated this French registry (XP and BT). DFG-project 441416480 (PL and DI). The ADAPT SFI Centre for Digital Content Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is cofunded under the European Regional Development Fund (ERDF) through Grant # 13/RC/2106_P2. JS is a Wellcome-IRish Clinical Academic Training (iCAT) Fellow, and this work was performed within the Irish Clinical Academic Training (iCAT) Programme, supported by the Wellcome Trust and the Health Research Board (Grant Number 203930/B16/2), the Health Service Executive, National Doctors Training and Planning and the Health and Social Care, Research and Development Division, Northern Ireland. German Research Foundation/Deutsche Forschungsgemeinschaft DFG-LA-1339/5-1 (PL).

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Ethics approval This study is part of a data reuse project. Ethical approval for the collection, sharing and analysis of patient data within the FAIRVASC consortium is governed locally at each registry site. Written informed consent was obtained directly from patients in five out of six participating registries. In the Skåne cohort, informed consent was not obtained and not required according to the ethical approval.

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.

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.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: https://creativecommons.org/licenses/by/4.0/.

ORCID iDs
Karl Gisslander http://orcid.org/0000-0003-1915-2656
Matthew Rutherford http://orcid.org/0000-0001-5626-2470
Louis Aslett http://orcid.org/0000-0003-2211-233X
Neil Basu http://orcid.org/0000-0003-4246-3145
François Dradin http://orcid.org/0009-0009-3377-0996
Lucy Hederman http://orcid.org/0000-0001-6073-4063
Zdenka Hruskova http://orcid.org/0000-0002-6635-365X
Hicham Kardaouli http://orcid.org/0000-0008-8403-0982
Peter Lamprecht http://orcid.org/0000-0001-5929-868X
Sabina Licholai http://orcid.org/0000-0001-7008-2492
Jacke Musial http://orcid.org/0000-0002-0955-1808
Declan O’Sullivan http://orcid.org/0000-0003-1990-3548
Xavier Puechal http://orcid.org/0000-0003-3573-9207
Jennifer Scott http://orcid.org/0000-0001-8837-5250
Märten Selmark http://orcid.org/0000-0002-1221-0772
Richard Straka http://orcid.org/0000-0007-2572-3941