



This is a repository copy of *Using routine primary care data in research: (in)efficient case studies and perspectives from the Asthma UK Centre for Applied Research*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/222017/>

Version: Published Version

Article:

Tibble, H. orcid.org/0000-0001-7169-4087, Alyami, R.A. orcid.org/0000-0001-6545-5346, Bush, A. et al. (15 more authors) (2025) Using routine primary care data in research: (in)efficient case studies and perspectives from the Asthma UK Centre for Applied Research. *BMJ Health & Care Informatics*, 32. e101134. ISSN 2632-1009

<https://doi.org/10.1136/bmjhci-2024-101134>

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial (CC BY-NC) licence. This licence allows you to remix, tweak, and build upon this work non-commercially, and any new works must also acknowledge the authors and be non-commercial. You don't have to license any derivative works on the same terms. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>



Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Using routine primary care data in research: (in)efficient case studies and perspectives from the Asthma UK Centre for Applied Research

Holly Tibble ¹, Rami A Alyami ^{2,3}, Andrew Bush,^{4,5} Steve Cunningham,⁶ Steven Julious,⁷ David Price,^{8,9} Jennifer K Quint,⁵ Stephen Turner,¹⁰ Kay Wang,¹¹ Andrew Wilson,¹² Gwyneth A Davies,¹³ Mome Mukherjee,¹⁴ Amy Hai Yan Chan,¹⁵ Deepa Varghese,¹⁴ Tracy Jackson,¹⁴ Noelle Morgan,¹⁶ Luke Daines,¹⁷ Hilary Pinnock¹⁸

To cite: Tibble H, Alyami RA, Bush A, *et al.* Using routine primary care data in research: (in)efficient case studies and perspectives from the Asthma UK Centre for Applied Research. *BMJ Health Care Inform* 2025;**32**:e101134. doi:10.1136/bmjhci-2024-101134

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjhci-2024-101134>).

Received 16 May 2024

Accepted 22 November 2024



© Author(s) (or their employer(s)) 2025. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

For numbered affiliations see end of article.

Correspondence to

Dr Holly Tibble;
htibble@ed.ac.uk

ABSTRACT

Aim We aimed to identify enablers and barriers of using primary care routine data for healthcare research, to formulate recommendations for improving efficiency in knowledge discovery.

Background Data recorded routinely in primary care can be used for estimating the impact of interventions provided within routine care for all people who are clinically eligible. Despite official promotion of ‘efficient trial designs’, anecdotally researchers in the Asthma UK Centre for Applied Research (AUKCAR) have encountered multiple barriers to accessing and using routine data.

Methods Using studies within the AUKCAR portfolio as exemplars, we captured limitations, barriers, successes, and strengths through correspondence and discussions with the principal investigators and project managers of the case studies.

Results We identified 14 studies (8 trials, 2 developmental studies and 4 observational studies). Investigators agreed that using routine primary care data potentially offered a convenient collection of data for effectiveness outcomes, health economic assessment and process evaluation in one data extraction. However, this advantage was overshadowed by time-consuming processes that were major barriers to conducting efficient research. Common themes were multiple layers of information governance approvals in addition to the ethics and local governance approvals required by all health service research; lack of standardisation so that local approvals required diverse paperwork and reached conflicting conclusions as to whether a study should be approved. Practical consequences included a trial that over-recruited by 20% in order to randomise 144 practices with all required permissions, and a 5-year delay in reporting a trial while retrospectively applied regulations were satisfied to allow data linkage.

Conclusions Overcoming the substantial barriers of using routine primary care data will require a streamlined governance process, standardised understanding/application of regulations and adequate National Health Service IT (Information Technology) capability. Without policy-driven prioritisation of these changes, the potential of this valuable resource will not be leveraged.

BACKGROUND

Randomised controlled trials (RCTs) estimate the efficacy of an intervention (including medicine, device, procedure or practice) delivered under controlled circumstances. While considered ‘gold standard’, RCTs recruit only a proportion of the potential population to which the intervention could be applied and are typically under-representative of underserved and minority populations.¹ In contrast, pragmatic trials are designed to estimate the effectiveness of the intervention in a routine clinical setting.² Increasingly, agencies are recognising the trade-off between the data precision of constrained RCTs and the enriched volume and reach of data that can be obtained by pragmatic designs. The US Food and Drug Administration, UK Medicines and Healthcare products Regulatory Agency and European Medicines Agency are now considering how to enable pragmatic designs to inform licensing of medicines (currently possible only for RCTs).³

Implementation research, which seeks to understand what works when an evidence-based intervention is provided as a routine service,⁴ typically uses pseudoanonymised data extracted from electronic health records (EHRs) to include all those clinically eligible (as opposed to recruited to research). In addition to experimental designs, observational studies can provide evidence of the impact of changes to clinical guidelines, policies or disease outbreaks. EHRs provide longitudinal patient data over large populations, including those with rare health conditions or specific demographic characteristics.⁵

Research conducted using EHRs has the potential to be time and cost efficient,⁶ and

ease the burden (both on the researcher and participant) of primary data collection while minimising the bias of overt observation and inaccurate recall.⁷ Anecdotally, however, these benefits are not always realised because of time-consuming technical and governance barriers to accessing primary care data.^{8,9}

Asthma UK Centre for Applied Research (AUKCAR) is a network of academics and partners, including people living with asthma, collaborating to improve care for people living with asthma. Meaningful collaboration with patients and public members is central to our work to ensure we are undertaking research that is of patient interest and benefit. Using AUKCAR portfolio studies as exemplars, we aimed to explore challenges in accessing primary care data and formulate recommendations to improve efficiency for EHRs in research.

METHODS

Case studies and data usage

We identified 13 UK-based, asthma-focused case studies conducted by AUKCAR investigators, which used routinely collected primary care EHR data in the study design (table 1). A detailed overview of all studies is presented in online supplemental appendix A.

The EHRs were used in three ways in these studies:

1. To identify eligible research participants. 10 studies used EHRs to inform recruitment in patient-level intervention studies and identify patients matching the inclusion criteria in observation studies. IMP²ART¹⁰ and TRAINS¹¹ recruited at the practice level, identifying their population of interest from the EHR. CHILL¹² recruited in schools and later linked to EHRs.
2. To ascertain exposures for study participants. 12 studies (all but TRAINS) used comprehensive medical histories recorded over extended periods. For example, the A4Sys study¹³ estimated asthma severity based on medication prescribed in primary care.
3. To collect outcome data. All the case studies used EHR data to assess outcomes (eg, incidence of asthma attacks).

Information synthesis

Data were captured (by HT) through a combination of extraction from published literature from the case studies, and interviews with the principal investigators, project managers, patient and public involvement contributors and data analysts. Individual hurdles and facilitators were coded (by HT in discussion with HP) from both literature and interviews from case studies, using grounded theory, an inductive and iterative form of thematic analysis. The code list was subsequently shared with all collaborators for further additions and clarifications. From this list, items were categorised into three overarching themes, which were then again agreed by all collaborators.

BARRIERS AND CHALLENGES

Three themes were identified related to barriers to efficient use of routine data for research in the case studies, with exemplars: information governance (table 2), recruitment of general practices (table 3) and data issues (table 4).

Information governance

While healthcare service providers have a fundamental duty to protect the confidentiality of the patients in their care, they equally have a duty to share information safely when it is of benefit to their care, as stated by the seventh Caldicott Principle.¹⁴ However, overly cautious interpretation of data protection rules, driven by fear of breaching confidentiality and cultural barriers, can be a substantial barrier to this duty, causing frustration among researchers and hindering important public health initiatives.

Research in National Health Service (NHS) practices, and using data derived from identifiable NHS records, requires several layers of regulatory approvals, research database approvals, NHS trust approvals and IT (Information Technology) provider approvals, in addition to the ethics and local research governance approvals applicable to all health service research. Governance approval is requested from the Health Research Authority (HRA) in England and Wales, Health and Social Care Northern Ireland (HSCNI) in Northern Ireland or NHS Research Scotland (NRS) in Scotland, through the Integrated Research Application System (IRAS) in parallel to the ethics application. In contrast to ethical approval which is provided once at the national level and accepted throughout the UK, Research and Development (R&D), capacity and capability (C&C) and information governance are conducted locally, following diverse and locally defined processes.

Diverse local governance in the IMP²ART cluster randomised implementation trial

IMP²ART practices are participants in a trial that uses routine data to evaluate the effectiveness of a strategy implementing supported self-management for asthma. The process is described as an exemplar IMP²ART practice in online supplemental appendix B and the key barriers are summarised below.

1. An interested general practitioner (GP) practice consulted their data protection officer, who might support the study or recommend that the practice not engage, due to their diverse interpretations of data protection legislation. An overcautious data protection officer could block participation in the study, despite multiple practices participating in other areas.
2. Before an interested practice could be recruited, regional R&D approvals were required. In England, this was C&C approval by NHS R&D Offices (of which there are over 400) in England. Each C&C required submission of documents attached to a series of emails (NHS email service does not accept large attachments or zip files) and the local R&D offices could not access

Table 1 Data sources in case studies

Study type	Short name	Brief study description	EHR uses:			Data sources linked to primary care
			Identify participants	Ascertain exposures	Collect outcomes	
Intervention Studies	PLEASANT	Investigation whether sending a postal letter for school-aged children, reminding them to renew their asthma medication prescriptions in time for the new school year start, improved prescription uptake.	Yes	Yes	Yes	None
	TRAINS	Investigating whether informing GP practices of the results of the PLEASANT trial motivated them to implement the intervention in their own practice.	No	No	Yes	None
	ARRISA	Investigating whether flagging high-risk asthma patients improved patient outcomes.	Yes	Yes	Yes	
	ARRISA-UK	Investigating whether flagging high-risk asthma patients, alongside a web-based training intervention, improved patient outcomes.	Yes	Yes	Yes	Secondary Care
	RAACENO	Investigating whether FeNO could provide an objective index to guide and stratify asthma treatment in children and improve patient outcomes.	Yes	Yes	Yes	None
	SPIROMAC	Investigating whether asthma treatment guided by spirometry plus symptoms (using the RAACENO algorithm), compared with symptoms alone, reduced asthma attack incidence.	Yes	Yes	Yes	
	IMP ² ART	Investigating whether targeted resources and training increased the provision of asthma action plans for self-management support.	No	Yes	Yes	None
	DEFINE	Investigating whether an online, primary care, FeNO-guided asthma management intervention reduces the risk of an acute asthma exacerbation compared with usual care.	Yes	Yes	Yes	None
Intervention development projects	ADxDA	Investigating whether a prediction model for asthma diagnosis in children and young people could be used as a clinical decision support system for use in primary care.	Yes	Yes	Yes	None
	A4Sys	Investigating whether a prediction model for asthma attack incidence in adults could be used as a clinical decision support system for use in primary care.	Yes	Yes	Yes	Secondary Care, Mortality

Continued

Table 1 Continued

Study type	Short name	Brief study description	EHR uses:			Data sources linked to primary care
			Identify participants	Ascertain exposures	Collect outcomes	
Observational studies of policy or practice change impact	SABINA	Investigating whether SABA use is associated with clinical outcomes in asthma.	Yes	Yes	Yes	Secondary Care
	CHILL	Investigating whether London's Ultra Low Emission Zone is associated with reduced air pollution in London and improved children's health outcomes.	No	Yes	Yes	Secondary Care
	EAVE II (asthma component)	Investigating whether there was a lower rate of asthma hospital admissions and deaths during the COVID-19 lockdown.	Yes	Yes	Yes	Secondary Care, Vaccinations, Virology, Mortality

EHR, Electronic Health Records; FeNO, fractional exhaled nitric oxide; GP, general practitioner; SABA, short-acting β_2 -agonist.

the files directly from the HRA repository. In Scotland, there was a similar process of local approvals from each of the 14 NHS Health Boards.

3. Regional IT provider approval was required, requiring further submission documents including a local Data Protection Impact Assessment and local checks would

be carried out to establish technical and governance compatibility with local practice computer systems.

Only once final approvals were in place could data extraction proceed, and the research team confirmed that practice was eligible for the trial and randomised. Moving through this process and achieving a data extraction for

Table 2 Information governance examples of challenges with using routine data in primary care research

Subtheme	Description	Exemplar
Inefficient processes	Research in NHS practices, and using NHS data, requires several layers of regulatory approvals, research database approvals, NHS trust approvals, and IT provider approvals, in addition to the ethics and local research governance approvals applicable to all health service research.	Highlighted by all study investigators in all the studies. An example of all steps required for the IMP ² ART study provided in Figure 1
Lack of standardised governance training	There is no training to ensure standardised interpretation of regulations, resulting in individuals creating and governing their own. Similarly, the researchers may be required to complete multiple virtually identical training courses from all involved institutions.	Researchers in the IMP ² ART and ARRISA-UK were each required to undertake multiple (very similar) governance training modules from multiple institutions, in order to receive, manage and analyse routine data.
Free-text data access	There is a wealth of patient data which is captured in free-text medical notes, and is never coded into structured fields. ¹⁹ Free-text data carry a substantially higher risk of confidentiality breach than coded data, and thus require further governance assurances.	The ADxDA study was unable to dependably ascertain data about presenting symptoms which were vital for reliable asthma diagnosis. ²⁰ The A4Sys study encountered great difficulty ascertaining when oral steroid prescriptions are prescribed for acute asthma attacks using coded data (rather than flare-ups of other immunological diseases, eg).
Additional steps for data linkage	Linkage between datasets may require the use of substantial identifiable data.	The ARRISA-UK study needed to collect identifiable data from patients' records to facilitate linkage between GP data and Hospital Episode Statistics, even though hashed NHS numbers were present in both datasets.
Changes and variations in governance processes	Governance approvals processes are subject to change.	For the ARRISA-UK study, a change to the NHS Digital governance processes during the study meant that previously randomised practices needed new approvals to conduct the approved linkage between primary care and secondary care data. Having completed the trial, many practices refused to undergo this additional process, and so were lost from subsequent analyses.
	Data protection officers and other responsible research staff had diverse interpretations of the legal requirements related to data access, resulting in a seemingly random element to approvals.	Highlighted by all the UK-wide studies not using existing research databases (eg, ARRISA, ARRISA-UK, RAACENO, IMP ² ART, SPIROMAC, DEFINE, CHILL and EAVE II)

NHS, National Health Service.

Table 3 General practice recruitment examples of challenges with using routine data in primary care research

Subtheme	Description	Exemplar
Individual practice recruitment for patient-level data	Primary care practices need to be recruited individually where patient-level data is to be extracted directly from the point of care. Researchers are often required to leverage several independently operating networks to access sufficient patients for their studies.	IMP ² ART study advertised through the NIHR Clinical Research Network (52% of recruited practices), NHS Research Scotland (7%), The Optimum Patient Care network (6%) and Education for Health newsletters (2%), as well as to practices already known to the IMP ² ART study team (3%), and through individual prospective identification (17%). The final 13% were recruited by word of mouth or other method.
Opt-in process for practices to join national anonymised data hubs	Individual practices must opt-in to inclusion in national research datasets, such as CPRD.	Typically, fewer than 20% of GPs in Scotland agree to share data with researchers. ²⁴ This is mirrored in England, where a whole-nation (54 million people) EHR resource enabled research on COVID-19 and cardiovascular disease, ²⁵ compared with CPRD which covers around 25% of the population.
Full nation data access	Full nation data is usually only available for NHS researchers.	In Scotland, the 2019 Joint Controller and Information Sharing Agreement between GP contractors and contracting Health Boards mean that all data from GP practices are automatically shared with their regional NHS Health Board and can be then transferred to a national database for auditing and NHS Scotland's use. Permissions were granted to the EAVE II study due to the special circumstances of the COVID-19 pandemic. ²⁶ However, once this permission expired, longer-term monitoring studies now required additional permissions—some of which rendered planned analyses infeasible.

EHR, electronic health record; GP, general practitioner; NHS, National Health Service.

an IMP²ART practice took a median of 10 weeks, but a third of the 144 participating practices were delayed by more than 6 months and 6% were delayed by over a year. 37 practices failed to reach the point of randomisation, meaning that 181 practices had to be recruited to achieve the sample size of 144, an over-recruitment of 20%. One practice withdrew after randomisation because a new data protection officer advised against the study contradicting the previous opinion.

Retrospective requirements in the ARRISA-UK cluster randomised implementation trial

Linkage of two datasets requires a linkage key, the common variable between the two sources which indicates that records should be linked together. In the ARRISA-UK study, linkage was planned between primary care and secondary care data, using the non-identifiable hashed NHS patient number. However, a change to the NHS Digital governance processes during the study meant that supplementary identifiable data were required for data linkage, to provide extra certainty that records were

accurately identified in order to provide reliable linkage with secondary care records in which no NHS number was recorded. As such, the 222 previously randomised practices needed new approvals to conduct the previously approved linkage between primary care and secondary care data as well as updating working to be consistent with the new General Data Protection Regulation (GDPR) related policies and regulation. Nearly 15% of practices did not agree to undergo this additional process and were therefore excluded from the analyses. The time required to design the revised system, obtain approvals and prepare and obtain data-sharing agreements resulted in an overall delay in reporting the trial of approximately three and a half years.

Recruitment of general practices

National research networks, such as the NIHR Clinical Research Networks (CRNs) in England, and the NRS Primary Care Network in Scotland support practice recruitment, but the process of accruals assumes and rewards individual patient recruitment. Studies recruiting

Table 4 Data-based challenges with using routine data in primary care research

Subtheme	Description	Exemplar
Inconsistent patient-based data	Some exposures may also be more likely to be documented in those with poorer health who may have more frequent contact with their healthcare providers, thus confounding the association between the exposure and certain poor clinical outcomes.	Individuals with severe asthma and propensity for attacks may have been more frequently self-testing for SARS-CoV-2 during the COVID-19 pandemic, which may have confounded the association between COVID-19 diagnosis and subsequent asthma attacks in this group of patients in the EAVE II study. ²⁷
Inconsistent research practices	There are not standardised and validated code lists and rule sets for ascertaining clinical features from primary care data.	The ADxDA study highlighted how various approaches to asthma diagnosis ascertainment may underestimate or overestimate the sample size. ²⁸
Rarely recorded data	Certain key confounders are often not possible to adjust for due to limited availability in EHRs, in particular tobacco (and vape) exposure and ethnicity. ^{28 29}	In the EAVE II study, the need to assess differences in vaccination, infection, treatment and outcomes by ethnic group was limited by lack of data in population datasets. ³⁰
Lack of negative information	There is often an absence of negative information, for example observing that a patient is not experiencing a specific symptom.	The ADxDA study needed to ascertain from coded data whether certain clinical features predictive of an asthma diagnosis were present, absent, or unknown. They made the assumption that the absence of a (coded) record of a characteristic or symptom may be less informative for common symptoms such as wheeze, breathlessness and allergy. ²⁸
Lack of healthy patients	When data entry is restricted to times of ill health, or scheduled appointments in those with ongoing health problems, there is a systematic deficit of data related to individuals with no health concerns.	In the A4Sys study, ¹³ the highest recorded peak flow measurement was used to reference longitudinal measurements. However, peak flow is often only measured in clinical practice at symptomatic appointments and will thus be systematically lower than average.
Administrative data not available in research data	Administrative information, such as mode of consultation, professional context of the clinician consulted, time of consultation, and whether it was an unscheduled consultation, is not routinely available in primary care research datasets.	This information would have been very beneficial for ascertainment of patients' care pathways, and for economic evaluations such as conducted in the PLEASANT ³¹ and ARRISA-UK trials. ³²

practices and using routine data do not fit with this model as no patients are recruited as there are when practices act as sites. This caused a specific problem for the IMP²ART trial where a randomised practice was asked to withdraw by a CRN in favour of a trial where they could act as a recruiting site.

Data issues

Healthcare data are collected for the purposes of recording the delivery of an episode of clinical care and the data structure is influenced by the diversity of clinical presentations, individual practitioner habits, local organisational conventions and funding contexts. The translation of events from primary care consultations to research datasets undergoes many steps, each of which may be subject to between-person variance in how they

are conducted, as highlighted in [figure 1](#). At the second stage in particular, clinical coding practices vary hugely between healthcare providers and are influenced by organisation factors such as the software used, whether coding is guided by a formulary and/or codes are added by a dedicated practice coder. In addition, clinical records are primarily created for use by the clinical service; secondary use is rarely the focus of coding decisions by clinical practitioners.

Validating the standardisation and harmonisation of real-world data for secondary purposes can be challenging, and variability between clinicians, practices, regions and across time periods challenges the reliability of real-world evidence.¹⁵

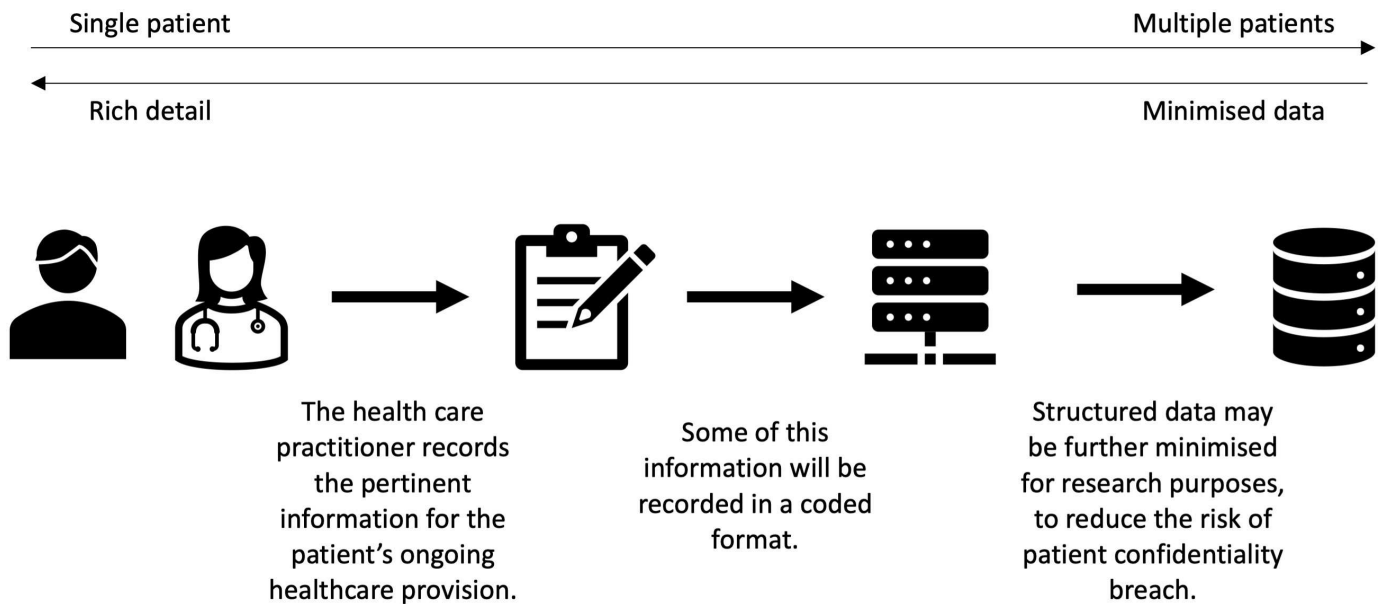


Figure 1 The translation of information from GP consultation to primary care research database. GP, general practitioner.

As well as inconsistencies in coding of clinical data, another issue highlighted in [table 2](#) was the limited availability of certain data. For example, lack of information on absent or non-reported symptoms, lack of data on healthy patients and limited administrative information in datasets available for research.

DISCUSSION

Summary of findings

The case studies that we have explored described both the opportunities and challenges of using routine data in applied and implementation research. A key strength was the convenience of collecting data for effectiveness outcomes, health economic assessment and process evaluation, in one data extraction. Crucially for implementation research, routine data can assess the impact of a new intervention on a whole population. Working with an established database could streamline processes of recruitment, ethics and regulatory approvals, data collection.

Despite recognising the benefits, many researchers described their frustration with barriers that demonstrably hindered research timelines, delayed analyses and were perceived to require many weeks to be spent on inefficient bureaucratic processes. Our case studies illustrate months of wasted researcher time, costly extensions and frustrated primary care practices blocked from participating in research. Inconsistent coding was a challenge for researchers, and the need to develop ways to extract data from free text was highlighted as a priority as was understanding the impact of external influences on coding practice. The delay in accessing data was often a surprise to our patients and public contributors, who expressed frustration and disappointment at the multiple barriers preventing access to valuable information that could improve people's lives.

Recommendations

We have collectively generated a list of five key recommendations to improve the efficiency of using primary care EHRs in research, as summarised in [table 5](#).

The first three recommendations relate to the current devolution of R&D, governance and IT approvals to local organisations. This is in marked contrast to ethics approval which is granted nationally and is not reviewed or challenged locally. The ethics application was seen as a time-consuming but reasonable process with agreed timelines that ensured timely completion. In the context of governance and IT approvals for UK-wide studies, the need for multiple local applications with diverse requests for paperwork, hugely varied timelines and reaching different conclusions amounted to a 'postcode lottery' and was universally frustrating to the research teams. Agreed standard paperwork, self-populating from the IRAS application where possible, should ensure that standardised information is available locally to inform decisions.

Variable interpretations by data protection (and other information governance) officers of the GDPR implications for use of routine data were a particular challenge highlighted by the researchers. Any uncertainty raised by a data protection officer will be likely to push decisions towards an overly risk-averse position suggesting that improved training and mentorship could support less experienced or overly cautious officers. There is a need to raise awareness of the potential of data-enabled trials in the UK and disseminate successful case studies² though knowledge of UK-wide positive approvals did not prevent individual data protection and local governance officers from disallowing willing practices from participating in the IMP²ART trial.

These issues were endorsed in a qualitative study by Mukherjee *et al*⁹ in the context of creating a learning

**Table 5** Summary of key recommendations to improve the efficiency of using routine data for primary-care based research

1	Process and timelines	Implement a streamlined process for regulatory approvals so that approvals proceed (where possible in parallel) and adhere to nationally agreed timelines.
2	Documentation	Standardise the variation in documentation requirements so that a single version of the Local Information Pack can be submitted centrally and be available to local R&D and IT departments.
3	Training and interpretation	Provide training and enforce local acceptance of nationally approved core documentation (eg, credentials of database organisations, standardised DPIAs) so that national approvals are accepted locally by default (as with ethical approvals) thus eliminating the current 'postcode lottery' in which local data protection officers, research staff, Trusts and practices apply different interpretation of regulations.
4	Prioritisation and capacity	Agree prioritisation of research and ensure capacity within local IT services to avoid delays in efficiently supporting research using routine data. Depending on context, this could apply at several levels; NHS Trusts, individual practices.
5	Methodological research	Allocate funding for methodological research into efficient use of routinely collected health data.

DPIA, Data Protection Impact Assessment; NHS, National Health Service; R&D, research and development.

health system, concluding that there needed to be a streamlined governance process akin to the NHS ethics system. Standardised training for trial sponsors, data protection officers, local governance and IT governance officers and all who are involved with decisions about research use of routine data, might enable a decision taken once to be understood and accepted through the UK. Inefficiencies in the current governance processes place undue pressure on the local governance officers, as tasks are repeated unnecessarily and paperwork reapproved in multiple locations leading to slow responses and delays to recruitment timelines. This then impacts on funders as additional costs are requested to extend projects to accommodate the delays. The costs and resources associated with over-recruitment of practices that ultimately cannot be randomised (20% in the IMP²ART trial) mean that extensions may not always be possible within the original budget. We did not capture the perspective of practice staff whose willingness to participate in a nationally approved study was blocked by a local decision (or lack of decision) but is unlikely to have been a positive experience.

Preconsent is receiving attention as a solution for some of the challenges in the context of commercial trials,¹⁶ which might streamline patient recruitment via EHR databases. The Clinical Practice Research Datalink (CPRD) provides a precedent for use of routine data in practice-level research: individual GP practices consent to sharing pseudoanonymised records, and patients are informed via a Fair Processing Notice displayed in their practice waiting room and on the practice website of their right to opt-out at any time.¹⁷ This enabled the PLEASANT¹⁸ and TRAINS¹¹ trials (table 1). Similarly, consent for low-risk and low-contact studies could be issued as an opt-out, with patients asked to opt-in to more sensitive research.

Clinical records are created to support the long-term provision of care, and decisions about suspected diagnosis and clinical management may be revised based on response to interventions and changes over time.

Meaningful ascertainment of clinical features thus requires a review of records in series and application of some heuristic processes. This is possible with routine data but can be challenging. The use of different coding systems in primary and secondary care (eg, SNOMED-CT and the International Classification of Diseases, ICD-10) presents a specific challenge as while there are cross-maps between terms in different systems, the nuances may be lost in translation. Changes in coding practices over time owing to external events such as the COVID-19 pandemic or changes to the Quality and Outcomes Framework (QOF) must be understood so that there is continued validity to the inferred meaning of codes. These influences external to an individual study, not only impact the practical logistics of the trial and interpretation of findings but may affect the risk–benefit considerations that underpin an ethical trial design or may disrupt the evidence generated.

Finally, there is great scope for methodological research which can fulfil the dual aims of improving utility in routine clinical practice while generating high-quality research-ready data. Given the diversity of primary care clinical practice and the vast range of codes available, the choice of any consultation can be overwhelming. Enabling practice formularies that limit the initial choice without restricting the final selection of a nuanced code could simplify the process for clinicians while standardising the coding for secondary use. Clinical records tell 'stories' recorded as free text to inform future consultations. Progress in the use of natural language processing can enable relevant clinical information to be extracted (with appropriate IG safeguards) from the unstructured data in EHRs, which may improve the detection of symptoms and improve the performance of clinical prediction models.^{19 20} Semiautomated processes for extracting codes from free text as it is entered would improve the validity of inferences from coded data as the clinician would be prompted to reject inappropriate suggestions. This is particularly pertinent in the rapidly changing

landscape of artificial intelligence; rapid evolution in governance guidelines is required to ensure that benefits can be achieved without compromising security.

Limitations and strengths

AUKCAR has links with the majority of UK applied researchers in asthma making it likely that we have included most large-scale research in this area. Together, we were able to comprehensively collate recurring issues preventing efficient research and provide detailed exemplars to describe the impact of these processes on publicly funded scientific research. Additionally, the AUKCAR collaboration includes researchers, healthcare providers and patients, and many members have multiple of these roles. As such, we have been able to thoroughly contemplate the conflicting needs and priorities of different parties.

This review voices the real-world experiences of the AUKCAR community throughout a series of large-scale studies, aiming to improve the lives of those living with asthma. The comprehensive 2022 review²¹ entitled 'Better, Broader, Safer: Using Health Data for Research and Analysis', by Goldacre and Morley presented 29 recommendations specifically related to information governance, as well as further recommendations in other domains. Many of these recommendations (particularly those under the subheading 'Enhanced usability for IG and ethics processes') are reinforced by the case studies we have described. For example, the seventh recommendation describes the objective to create a national centre for governance and regulations, which could develop standardised documentation and training, and provide top-level insights which would facilitate processes. In our current landscape, many people given 'expert' IG roles feel unprepared and understandably err on the side of caution.

Our study is limited to the UK research context (particularly pertinent, given the longevity of primary care EHR databases in the UK) though international literature has also highlighted similar issues,^{22 23} and the enablers and barriers identified have potential relevance to other countries using routinely collected data. Healthcare systems at an earlier stage of development of similar resources may be able to optimise the benefits and avoid the barriers that challenge our UK research. Similarly, we focused on asthma as an exemplar, but it is unlikely that experiences will be different in other disease areas (apart from COVID-19 when many of the regulatory processes were temporarily eased to enable a prompt response to the pandemic).

Finally, we note that while using routinely collected primary care data broadens the population compared with studies which specifically recruit participants, there are still some people not registered with, or not attending, a primary care practice, who will be missed in the population denominator. Linkage to other national or regional datasets may help detect the non-registered,

but those who do not consult about their condition will still be missed.

CONCLUSIONS

Routinely collected EHRs provide opportunities for primary care research but there are significant barriers to overcome will require a commitment across a range of stakeholders including prioritisation by policy-makers, development of streamlined governance processes, commitment to standardised training, adequate NHS IT capability and methodological research if the potential of this valuable resource is to be leveraged.

Author affiliations

- ¹The University of Edinburgh Usher Institute of Population Health Sciences and Informatics, Edinburgh, UK
- ²University of Sheffield, Sheffield, Saudi Arabia
- ³Respiratory Therapy Department, Jazan University, Jazan, Saudi Arabia
- ⁴National Heart and Lung Institute, London, UK
- ⁵Imperial College London, London, UK
- ⁶Department of Child Life and Health, University of Edinburgh, Edinburgh, UK
- ⁷University of Sheffield, Sheffield, UK
- ⁸Observational and Pragmatic Research Institute, Singapore
- ⁹University of Aberdeen, Aberdeen, UK
- ¹⁰Child Health, University of Aberdeen, Aberdeen, UK
- ¹¹University of Oxford, Oxford, UK
- ¹²University of East Anglia, Norwich, UK
- ¹³Swansea University Medical School, Swansea, UK
- ¹⁴The University of Edinburgh, Edinburgh, UK
- ¹⁵The University of Auckland, Auckland, New Zealand
- ¹⁶Patient Contributor, Edinburgh, UK
- ¹⁷Centre for Medical Informatics, The University of Edinburgh Usher Institute of Population Health Sciences and Informatics, Edinburgh, UK
- ¹⁸College of Medicine and Veterinary Medic, The University of Edinburgh Usher Institute of Population Health Sciences and Informatics, Edinburgh, UK

X Holly Tibble @hollytibble and Rami A Alyami @RamiAlyami

Acknowledgements We would like to additionally thank Victoria Hammersley, Natalia Reglinska-Matveyev, Francis Appiagyei, Colin Simpson, and Mousumi Sengupta for their contributions to this paper.

Contributors This idea for this study was primarily conceived by HP. HT collated the evidence from coauthors via face-to-face and email correspondence and was the lead author, corresponding author and guarantor. Authors RAA, AB, SC, LD, SJ, DP, JKQ, ST, KW, AW, GAD, MM and AHYC input to the design of the paper, contributed data on their own studies and provided edits to the final draft. Authors DV, TJ and NM added insights from their own experiences, and feedback on the general structure of the paper. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. The corresponding author has the right to grant on behalf of all authors and does grant on behalf of all authors, a worldwide licence to the publishers and its licensees in perpetuity, in all forms, formats and media (whether known now or created in the future), to (1) publish, reproduce, distribute, display and store the contribution, (2) translate the contribution into other languages, create adaptations, reprints, include within collections and create summaries, extracts and/or abstracts of the contribution, (3) create any other derivative work(s) based on the contribution, (4) to exploit all subsidiary rights in the contribution, (5) the inclusion of electronic links from the contribution to third party material wherever it may be located and (6) licence any third party to do any or all of the above.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

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 Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Holly Tibble <http://orcid.org/0000-0001-7169-4087>

Rami A Alyami <http://orcid.org/0000-0001-6545-5346>

REFERENCES

- Brown T, Jones T, Gove K, *et al*. Randomised controlled trials in severe asthma: selection by phenotype or stereotype. *Eur Respir J* 2018;52:1801444.
- Sydes MR, Barbachano Y, Bowman L, *et al*. Realising the full potential of data-enabled trials in the UK: a call for action. *BMJ Open* 2021;11:e043906.
- Baumfeld Andre E, Reynolds R, Caubel P, *et al*. Trial designs using real-world data: The changing landscape of the regulatory approval process. *Pharmacoepidemiol Drug Saf* 2020;29:1201–12.
- Peters DH, Adam T, Alonge O, *et al*. Implementation research: what it is and how to do it. *BMJ* 2013;347:f6753bmj.f6753.
- Glicksberg BS, Johnson KW, Dudley JT. The next generation of precision medicine: observational studies, electronic health records, biobanks and continuous monitoring. *Hum Mol Genet* 2018;27:R56–62.
- Treweek S, Altman DG, Bower P, *et al*. Making randomised trials more efficient: report of the first meeting to discuss the Trial Forge platform. *Trials* 2015;16:261.
- Roche N, Anzueto A, Bosnic Anticevich S, *et al*. The importance of real-life research in respiratory medicine: manifesto of the Respiratory Effectiveness Group: Endorsed by the International Primary Care Respiratory Group and the World Allergy Organization. *Eur Respir J* 2019;54:1901511.
- Powell GA, Bonnett LJ, Tudur-Smith C, *et al*. Using routinely recorded data in the UK to assess outcomes in a randomised controlled trial: The Trials of Access. *Trials* 2017;18:389.
- Mukherjee M, Cresswell K, Sheikh A. Identifying strategies to overcome roadblocks to utilising near real-time healthcare and administrative data to create a Scotland-wide learning health system. *Health Informatics J* 2021;27:1460458220977579.
- McClatchey K, Hammersley V, Steed L, *et al*. IMpLementing IMProved Asthma self-management as RouTine (IMP²ART) in primary care: study protocol for a cluster randomised controlled implementation trial. *Trials* 2023;24:252.
- Alyami RA, Simpson R, Oliver P, *et al*. TRial to Assess Implementation of New research in a primary care Setting (TRAINS): study protocol for a pragmatic cluster randomised controlled trial of an educational intervention to promote asthma prescription uptake in general practitioner practices. *Trials* 2022;23:947.
- Tsocheva I, Scales J, Dove R, *et al*. Investigating the impact of London's ultra low emission zone on children's health: children's health in London and Luton (CHILL) protocol for a prospective parallel cohort study. *BMC Pediatr* 2023;23:556.
- Tibble H, Tsanas A, Horne E, *et al*. Predicting asthma attacks in primary care: protocol for developing a machine learning-based prediction model. *BMJ Open* 2019;9:e028375.
- The Caldicott Committee. Report on the review of patient-identifiable information. 1997.
- Abbaszanjani H, Torabi F, Bedston S, *et al*. Harmonising electronic health records for reproducible research: challenges, solutions and recommendations from a UK-wide COVID-19 research collaboration. *BMC Med Inform Decis Mak* 2023;23:8.
- O'Shaughnessy J. Commercial clinical trials in the uk: the lord o'Shaughnessy review. UK Department of Health & Social Care; 2023.
- Quint JK, Moore E, Lewis A, *et al*. Recruitment of patients with Chronic Obstructive Pulmonary Disease (COPD) from the Clinical Practice Research Datalink (CPRD) for research. *NPJ Prim Care Respir Med* 2018;28:21.
- Horspool MJ, Julious SA, Boote J, *et al*. Preventing and lessening exacerbations of asthma in school-age children associated with a new term (PLEASANT): study protocol for a cluster randomised control trial. *Trials* 2013;14:297.
- Ford E, Curlewis K, Squires E, *et al*. The Potential of Research Drawing on Clinical Free Text to Bring Benefits to Patients in the United Kingdom: A Systematic Review of the Literature. *Front Digit Health* 2021;3:606599.
- Daines L, Bonnett LJ, Boyd A, *et al*. Protocol for the derivation and validation of a clinical prediction model to support the diagnosis of asthma in children and young people in primary care. *Wellcome Open Res* 2020;5:50.
- Goldacre B, Morley J. Better, broader, safer: using health data for research and analysis. 2022. Available: <https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis/better-broader-safer-using-health-data-for-research-and-analysis>
- Holmes JH, Beinlich J, Boland MR, *et al*. Why Is the Electronic Health Record So Challenging for Research and Clinical Care? *Methods Inf Med* 2021;60:32–48.
- O'Brien EC, Raman SR, Ellis A, *et al*. The use of electronic health records for recruitment in clinical trials: a mixed methods analysis of the Harmony Outcomes Electronic Health Record Ancillary Study. *Trials* 2021;22:465.
- Harron KL. Linking data to build a bigger picture of paediatric referral pathways. *Arch Dis Child* 2023;108:245–6.
- Wood A, Denholm R, Hollings S, *et al*. Linked electronic health records for research on a nationwide cohort of more than 54 million people in England: data resource. *BMJ* 2021;373:826.
- Simpson CR, Robertson C, Vasileiou E, *et al*. Early Pandemic Evaluation and Enhanced Surveillance of COVID-19 (EAVE II): protocol for an observational study using linked Scottish national data. *BMJ Open* 2020;10:e039097.
- Davies GA, Alsallakh MA, Sivakumaran S, *et al*. Impact of COVID-19 lockdown on emergency asthma admissions and deaths: national interrupted time series analyses for Scotland and Wales. *Thorax* 2021;76:867–73.
- Daines L, Bonnett LJ, Tibble H, *et al*. Deriving and validating an asthma diagnosis prediction model for children and young people in primary care. *Wellcome Open Res* 2023;8:195.
- Shi T, Pan J, Vasileiou E, *et al*. Risk of serious COVID-19 outcomes among adults with asthma in Scotland: a national incident cohort study. *Lancet Respir Med* 2022;10:347–54.
- Pareek M, Bangash MN, Pareek N, *et al*. Ethnicity and COVID-19: an urgent public health research priority. *The Lancet* 2020;395:1421–2.
- Julious SA, Horspool MJ, Davis S, *et al*. Open-label, cluster randomised controlled trial and economic evaluation of a brief letter from a GP on unscheduled medical contacts associated with the start of the school year: the PLEASANT trial. *BMJ Open* 2018;8:e017367.
- Smith JR, Noble MJ, Musgrave S, *et al*. The at-risk registers in severe asthma (ARRISA) study: a cluster-randomised controlled trial examining effectiveness and costs in primary care. *Thorax* 2012;67:1052–60.