Systematic approach to outcome assessment from coded electronic healthcare records in the DaRe2THINK NHS-embedded randomised trial

Wang, Xiaoxia; Mobley, Alastair; Tica, Otilia; Okoth, Kelvin; Nirantharakumar, Krishnarajah; Shukla, David; Kotecha, Dipak

DOI:
10.1093/ehjdh/ztac046

ESC European Society of Cardiology

European Heart Journal - Digital Health (2022) **3**, 426–436
https://doi.org/10.1093/ehjdh/ztac046

ORIGINAL ARTICLE

Systematic approach to outcome assessment from coded electronic healthcare records in the DaRe2THINK NHS-embedded randomized trial

Xiaoxia Wang [1,2], Alastair R. Mobley [1,2], Otilia Tica [1], Kelvin Okoth [3], Rebecca E. Ghosh[4], Puja Myles[4], Tim Williams[4], Sandra Haynes[5], Krishnarajah Nirantharakumar [3], David Shukla [3,6], and Dipak Kotecha [1,2,*], on behalf of the DaRe2THINK Trial Committees[†]

[1]Institute of Cardiovascular Sciences, University of Birmingham, Birmingham, UK; [2]Health Data Research UK Midlands, Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK; [3]Institute of Applied Health Sciences, University of Birmingham, Birmingham, UK; [4]Clinical Practice Research Datalink, Medicines and Healthcare products Regulatory Agency, London, UK; [5]Patient and Public Involvement team, Birmingham, UK; and [6]Primary Care Clinical Research, NIHR Clinical Research Network West Midlands, Birmingham, UK

Received 30 May 2022; revised 15 August 2022; online publish-ahead-of-print 16 September 2022

Aims Improving the efficiency of clinical trials is key to their continued importance in directing evidence-based patient care. Digital innovations, in particular the use of electronic healthcare records (EHRs), allow for large-scale screening and follow up of participants. However, it is critical these developments are accompanied by robust and transparent methods that can support high-quality and high clinical value research.

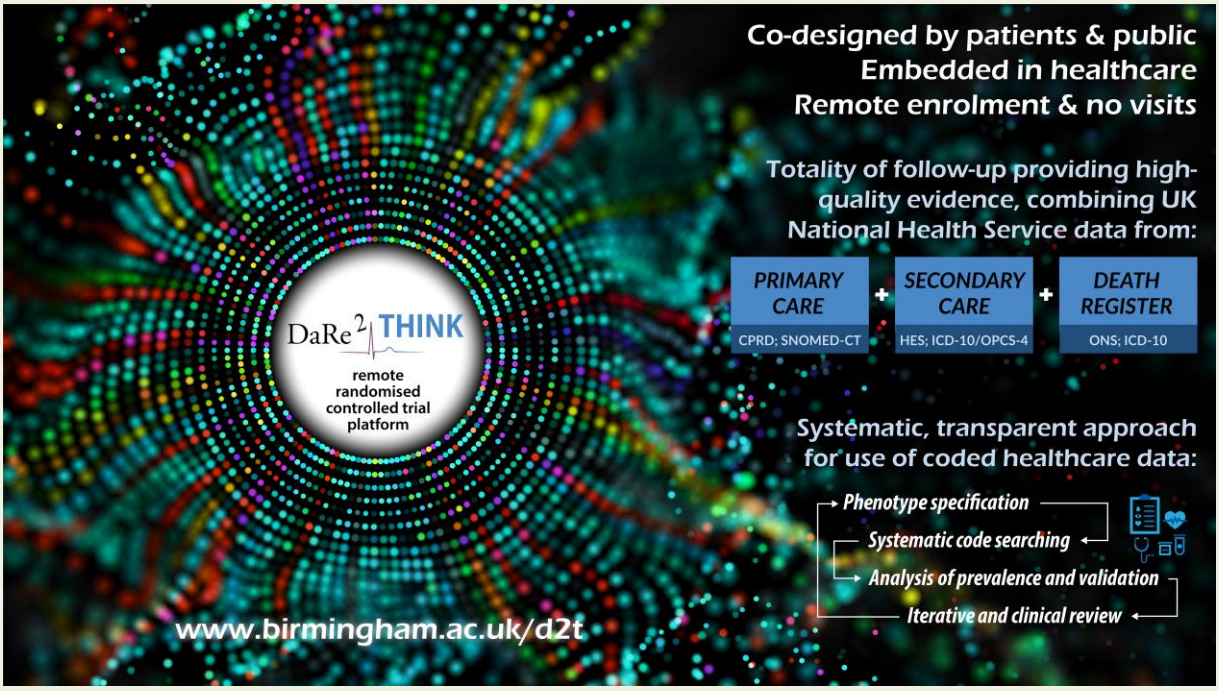
Methods and results The DaRe2THINK trial includes a series of novel processes, including nationwide pseudonymized pre screening of the primary-care EHR across England, digital enrolment, remote e-consent, and 'no-visit' follow up by linking all primary- and secondary-care health data with patient-reported outcomes. DaRe2THINK is a pragmatic, healthcare-embedded randomized trial testing whether earlier use of direct oral anticoagulants in patients with prior or current atrial fibrillation can prevent thromboembolic events and cognitive decline (www.birmingham.ac.uk/dare2think). This study outlines the systematic approach and methodology employed to define patient information and outcome events. This includes transparency on all medical code lists and phenotypes used in the trial across a variety of national data sources, including Clinical Practice Research Datalink Aurum (primary care), Hospital Episode Statistics (secondary care), and the Office for National Statistics (mortality).

Conclusion Co-designed by a patient and public involvement team, DaRe2THINK presents an opportunity to transform the approach to randomized trials in the setting of routine healthcare, providing high-quality evidence generation in populations representative of the community at risk.

* Corresponding author. Heritage Building, Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Mindelsohn Way, Birmingham, B15 2GW, UK. Tel: +44 121 3718122, Email: d.kotecha@bham.ac.uk

[†]See Acknowledgements.

Graphical Abstract



Keywords Randomized controlled trial • Electronic healthcare record • Coding • Atrial fibrillation • Anticoagulation • Primary care • Secondary care

Introduction

Rapid uptake of electronic healthcare record (EHR) systems across the world have led to the opportunity for large-scale, real-world, longitudinal clinical research.[1] In countries with national health systems, there is also the potential to link all EHRs for an individual patient, across both primary and secondary care, and from birth to death. Standardized coding systems provide the basis for harnessing medical information from different healthcare providers, with linkage allowing for a complete picture of each patient's history, healthcare utilization, and adverse events.[2] This is an attractive (and cost-efficient) prospect for clinical research, particular in countries such as the UK where the quality of EHR data is typically high due to the need for accurate assessment of healthcare utilization within the National Health Service (NHS).[3–5]

Observational clinical research has benefited for many decades from EHR data. Evolving technology and coding systems, coupled with more complete EHR coverage, have provided the opportunity for randomized controlled trials (RCTs) to be embedded alongside EHR systems, taking advantage of existing clinical data and follow up. The coronavirus pandemic has epitomized how innovations in the design and running of RCTs are critical to address unmet clinical need. However, RCTs remain too costly, tend to recruit selective patients within high-performing centres, and frequently do not match the real population at risk in our communities.

The DaRe2 approach (healthcare Data for pragmatic clinical Research in the NHS—primary 2 secondary) was designed to incorporate recent innovations in EHR and clinical research, combined with advances in mobile technology, to provide an end-to-end framework for RCTs embedded in Primary Care in England. The DaRe2THINK trial provides an exemplar of a 'remote-RCT' to address a key public health concern, requiring no physical patient visits despite being a clinical trial of an investigational medicinal product.[6] In this study, we outline the design features of DaRe2THINK, and the systematic approach used to define patient information and outcome events that will be accrued during the trial. All medical code lists and phenotypes are presented for full transparency, and for future use by other researchers using EHR data.

The DaRe2THINK trial

Using the digital potential of the EHR, DaRe2THINK is transformational project that will underpin a new trajectory for NHS-based research for patient benefit. DaRe2THINK is led by the University of Birmingham, in collaboration with the Clinical Practice Research Datalink (CPRD; part of the Medicines and Healthcare products Regulatory Agency). The trial is supported by the University Hospitals Birmingham NHS Foundation Trust, the National Institute for Health and Care Research (NIHR) Clinical Research

Network (West Midlands Primary Care), and Health Data Research UK Midlands.

DaRe2 platform features

(1) Automated secure screening of inclusion and exclusion criteria across >13 million NHS patients, including around one in four primary-care sites in England that are part of CPRD. This provides rapid and cost-efficient screening of a diverse and representative proportion of the UK population,[7] yet maintains patient privacy using pseudonymized records that can only be re-identified by local NHS staff.
(2) Access to national live data on the numbers of potentially eligible patients allows for smart adaptation of trial inclusion and exclusion criteria.
(3) Targeted enrolment of primary-care sites identified as having potentially eligible patients, streamlining recruitment by focusing resources on high-value sites.
(4) Dynamic adaptive screening of the population at each site, with updates on a weekly basis of newly eligible patients, reducing burden on local NHS staff and simplifying recruitment.
(5) Easy enrolment of patients without need for physical attendance, with remote e-consent using their own smartphone/tablet/computer, one-click randomization by primary-care staff, and drug prescription via usual clinical systems.
(6) 'No-visit' follow up, utilizing NHS records linked across national primary and secondary care for capture of endpoints without the patient attending the healthcare site, and no need for NHS investigators to complete case report forms.
(7) Regular scheduled patient-reported outcomes, with requests sent by automated text message to each participant's telephone and email, and secure data acquisition through online completion.
(8) Automated collation of safety outcomes from the primary-care NHS record, avoiding the possibility of missed events unknown to the local investigator.
(9) Co-design and management of all processes by a patient and public involvement (PPI) team,[8] supporting a patient-centric approach with sustainable and valued output.

DaRe2THINK design

DaRe2THINK will be the first exemplar of this system, and is focused on the intersection of key national priorities for health and social care services—the common heart rhythm disorder called atrial fibrillation (AF) and the impact this condition has on thromboembolic events, including long-term cognitive decline and vascular dementia. Patients with AF suffer from a high rate of morbidity, with one in four stroke patients having AF as a potential cause, half of AF patients developing heart failure that responds poorly to treatment, and the rate of death doubled at all ages.[9–11] Most patients with AF develop progressive subclinical cerebral damage over time,[12,13] with high rates of cognitive impairment and a 40% increased risk of dementia compared with normal sinus rhythm.[14,15] As the prevalence of AF is expected to double in the coming decades,[16] this will place an unsustainable burden on healthcare and social services.

DaRe2THINK is a pragmatic, NHS-embedded, open-label, event-driven, parallel-group RCT. It will test whether earlier use of anticoagulants is effective and cost-effective at preventing thromboembolic events (primary outcome) and cognitive decline (key secondary outcome) in patients with AF at low or intermediate risk of stroke. Around 3000 patients will be randomized to either starting any direct oral anticoagulant (used due to their excellent safety and efficacy profile[17,18]), or continuing standard-of-care where anticoagulation is instituted when the patient is older with clear risk factors for stroke.[19] Follow up includes yearly collection of cognitive and outcome data for 5 years. Consent also includes remote follow up at 10 years and then lifetime EHR outcomes for long-term conditions such as vascular dementia. A summary of the DaRe2THINK trial is presented in *Figure 1*. Due to the ability to access pseudonymized NHS records, DaRe2THINK will be able to adapt inclusion and exclusion criteria to reflect current changes in practice and extraneous factors (e.g. the coronavirus pandemic). Recruitment has commenced across England (https://www.birmingham.ac.uk/dare2think).

Patient and public involvement has been central to the development of the DaRe2THINK trial process from conception through to delivery. DaRe2THINK follows the PPI-POSITIVE approach for implementation of genuine public engagement at every stage of the trial, including patient contact, study management, analysis, and dissemination.[8] A plain English summary written by the PPI team is provided in *Table 1*.

Systematic coding methodology

The DaRe2 approach links and combines NHS data from primary-care, secondary-care, and other national databases in order to provide a full and complete picture of each participant's health and healthcare utilization. The UK public health system is free at the point of delivery to all citizens, with indirect reimbursement via local health authorities based on coded data for diseases and procedures performed. In this section, we detail the processes employed in the DaRe2THINK trial to systematically catalogue and incorporate all relevant codes, including for selection criteria, baseline variables, outcome endpoints. In addition, DaRe2THINK includes monthly automated searches of the primary-care record to assess for safety events. For transparency, all codes used are presented in Supplementary material online, *Appendix S1* for other researchers to see, comment, update, and re-use. The current trial protocol, including details on trial processes and sample size calculations, is available in Supplementary material online, *Appendix S2*. The DaRe2THINK trial adheres to the CODE-EHR best practice framework for the use of structured EHRs in clinical research.[2] This study meets all five of the CODE-EHR minimum standards, and will in addition meet all five standards for preferred criteria once completed; further details are presented in Supplementary material online, *Appendix S3*. The DaRe2THINK investigators are committed to the FAIR principles (Findable, Accessible, Interoperable, and Reusable).[20]

Data sources

Primary-care data are obtained through CPRD Aurum, a prospectively collected, population-based, pseudonymized medical record database that collects daily information directly from NHS primary-care sites that are part of the CPRD network and use the Egton

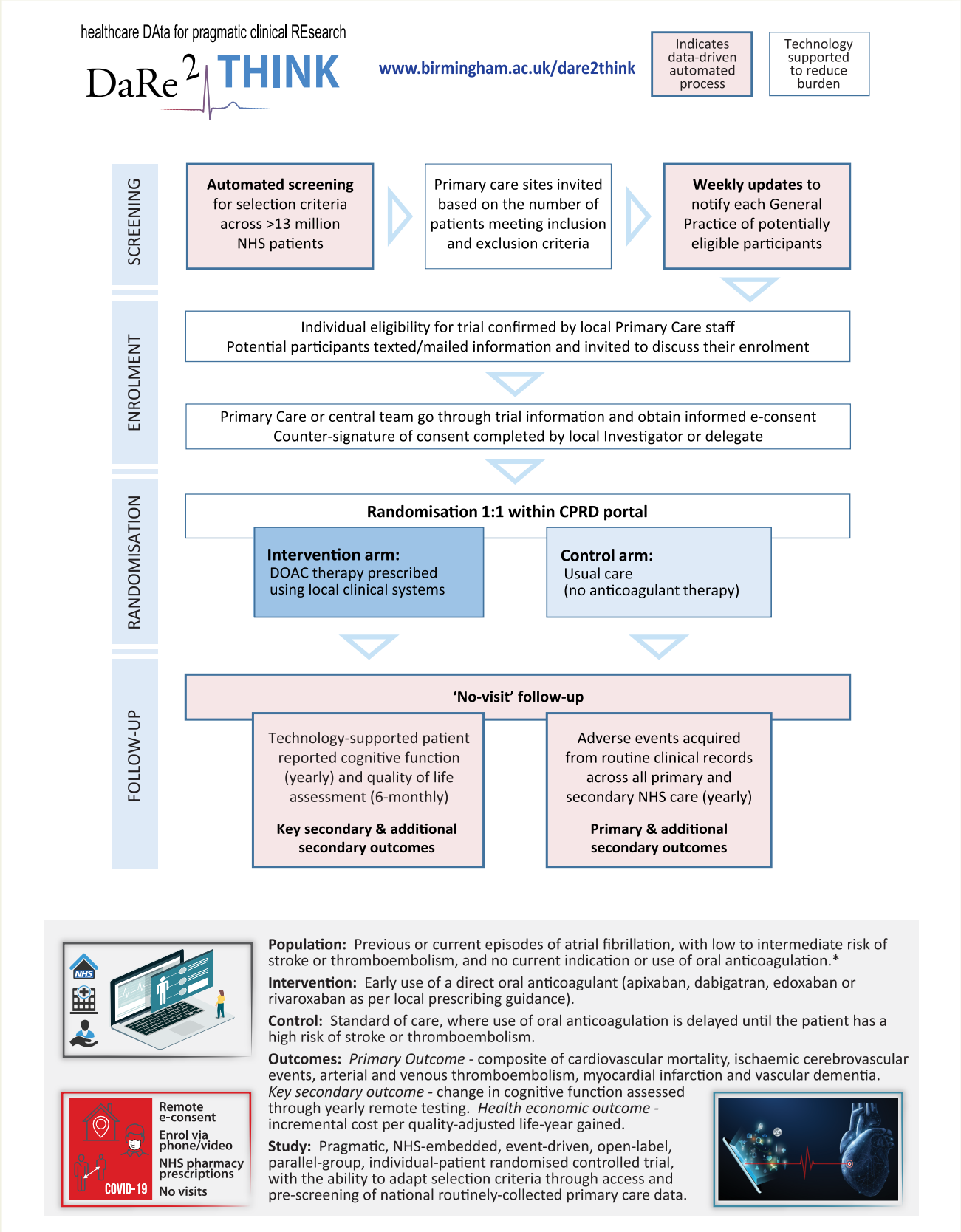


Figure 1 DaRe2THINK key innovations. Flow chart of the DaRe2 approach (healthcare data for pragmatic clinical research in the NHS—primary 2 secondary) and summary of the DaRe2THINK clinical trial. *Please see current protocol at www.birmingham.ac.uk/dare2think for full details of inclusion and exclusion criteria. CPRD, Clinical Practice Research Datalink; DOAC, direct oral anticoagulant; NHS, National Health Service.

Table 1 Plain English summary of DaRe2THINK

Challenges and opportunities	The NHS is unlike any other in the world, caring for people throughout their lives both in the community and in hospitals. At the heart of DaRe2THINK is that health data collected within these services can be used for the benefit of patients. Clinical trials are an important way to understand how new treatments can be used in the NHS, but many trials struggle to find the right patients, or be relevant to their needs. The DaRe2 approach (healthcare Data for pragmatic clinical Research in the NHS—primary 2 secondary) will test a new way of running trials based at general practitioner (GP) surgeries using routine NHS information. We will include patients that do not normally take part in clinical trials and follow them up without the need to revisit their GP or attend hospital. This approach could improve the health and well-being of those treated by the NHS, while reducing the time needed from staff and patients to engage in important research.
Addressing a gap in current healthcare	As an example of this new system, DaRe2THINK will target an issue of huge importance to patients, our NHS and the social care system. AF is a common heart rhythm condition that leads to a high chance of stroke, frequent hospital admissions, and poor quality of life. Patients also have a much higher risk of cognitive decline (trouble remembering, concentrating, or making everyday decisions) and dementia. This may be due to silent 'micro-strokes' that gradually damage the brain over time. Blood thinning tablets (anticoagulants) greatly reduce the number of patients with AF that will suffer a stroke, but are usually only given to older patients or those with other health issues. This may be too late to avoid dementia. It also leaves those younger than 65 years, and some patients aged 65–75 without treatment that could prevent these devastating complications.
Aim of DaRe2THINK	A new class of blood thinning tablets are now widely used in the NHS which are more convenient for patients to take, and have a lower risk of bleeding than older treatments. These drugs could provide an effective way to prevent strokes, brain damage, and dementia in later life for a broader group of patients, but this needs to be tested in a clinical trial.
How the trial works	With the support of a PPI team and a national network of research nurses and GPs, the trial will include 3000 patients from up to 600 GP surgeries across England. Each patient will either continue their current treatment or start an additional blood thinning tablet on a random basis. Patients will be followed up automatically within the NHS to look at the difference in those who suffer from strokes, blood clots, heart attacks, other problems with the blood vessels, and dementia. Patients will self-report their memory, reaction times, and quality of life using simple questionnaires through their mobile phone or the internet, again without needing to revisit their doctor.
Results and impact	DaRe2THINK will answer important questions for a growing number of patients with AF. The combination of information from the community as well as hospitals across the NHS will allow us to see whether these blood thinning tablets should be prescribed more widely. DaRe2THINK will allow us to develop and improve this new clinical trial system so that future research in the NHS will continue to benefit those patients most in need.

Medical Information Systems software system. As of March 2021 when DaRe2THINK was initiated, CPRD Aurum included 39 555 354 research acceptable patients of which 13 299 826 were actively registered (19.9% of the UK population) across 1375 active primary-care sites (15.3% of UK general practices). All baseline data for DaRe2THINK trial participants is extracted from CPRD Aurum, obviating the need for investigators to fill in case report forms. CPRD Aurum includes codes for diagnosis and non-prescription data (medcodeid), which is cross-mapped to SNOMED CT (UK edition)[21] and Read Version 2,[22] with prescriptions by product code (prodcodeid) mapped to the Dictionary of Medicines and Devices.[23]

Secondary care Hospital Episode Statistics (HES) admitted patient care data, which contains information on all admissions to NHS hospitals in England, as well as NHS-funded care at independent providers, are linked at the patient level to CPRD Aurum by NHS Digital. With around 99% of hospital activity in England funded by the NHS, this provides almost total coverage of healthcare utilization for the DaRe2THINK trial, including admission and discharge date, admission type (pre planned or emergency), primary diagnosis (reason for hospitalization), secondary diagnoses, and any procedures performed during the hospital stay. HES uses the International Classification of Diseases version 10 (ICD-10) for diagnosis codes,[24] and the Office of Population Censuses and Surveys Classification of Surgical Operation and Procedures (OPCS) version 4 for procedures.[25]

Death and cause of death are obtained via linkage with the Office for National Statistics (ONS) mortality database,[26] which includes all deaths in England occurring both within and outside any healthcare setting. Data in ONS are coded using ICD-10.

Development of the medical code lists

To achieve a comprehensive set of contemporary codes which are suitable for defining outcome and safety events in a clinical trial, we designed and employed a four-phase systematic framework to develop code lists for data extraction from the listed data sources. The approach is summarized in *Figure 2*. Throughout the four phases, we used DExtER, an automated platform for clinical epidemiology that was developed to support clinical code list development and validation of phenotypes.[28]

Phase 1: phenotype specification

Prior to searching for relevant medical codes, the DaRe2THINK trial selection criteria and outcome measures were first transformed into relevant phenotypes for each data source. This was an iterative process that accounted for the limitations of the data source and, where

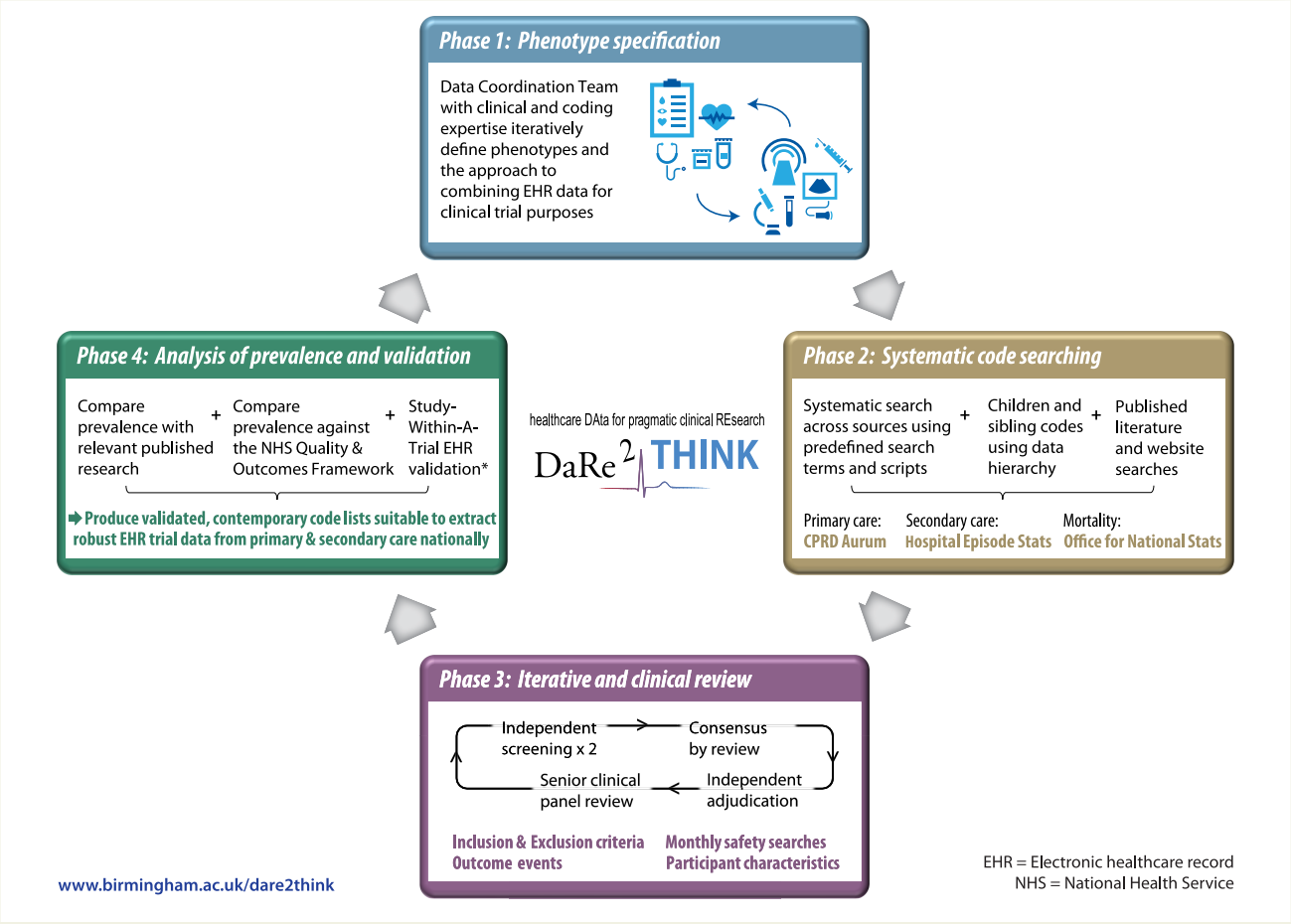


Figure 2 DaRe2THINK coding framework. Systematic, iterative approach to defining code lists from national healthcare data sources. *Study-Within-A-Trial to be performed in a subset of patients with clinical events identified using primary and secondary healthcare records.[27]

needed, combined phenotypes to derive a particular outcome. For example, 'major bleeding' is not recorded as such in any primary-care data set; hence, we combined a phenotype list for bleeding events with a list generated to identify concurrent hospitalization. This provided a code list for the safety outcome of 'bleeding resulting in hospitalization' based solely on CPRD Aurum data. This process was facilitated by the Data Coordination Team, a specific DaRe2THINK committee which includes clinical expertise in primary and secondary care, data scientists with knowledge and experience in using medical codes, and representatives of CPRD.

Phase 2: systematic code searching

A systematic search was carried out against the agreed phenotypes using the medical code dictionaries (SNOMED CT and Read codes for CPRD Aurum, and ICD-10 codes for HES and ONS). When searching medical code dictionaries, search terms and acronyms were agreed with clinical input. The hierarchy structure of each dictionary was used to find additional relevant children or sibling codes, and these were collated into a dynamic code list. The list was supplemented by any codes identified by other researchers after screening of previously published and relevant research. To facilitate future use, or if there are updates to coding systems during the trial follow-up period, all search terms and searching scripts are saved as meta data. Medication codes required in the phenotyping algorithms were extracted from the CPRD Aurum product lists based on both British National Formulary codes and drug substance names. CPRD provided established code lists for demographics and measurements of ethnicity, height, weight, blood pressure, etc.

Phase 3: iterative and clinical review

The codes found in the systematic search went through a multi-stage review process. This included review by data and clinical scientists. Code lists were reviewed independently by two individuals, marking codes for inclusion, exclusion, or further review. Discrepancies between the reviewers and any queries were resolved by consensus agreement after adjudication by a senior investigator. All resultant code lists were then further reviewed by senior clinical experts to ensure that code lists were correct and consistent.

Phase 4: analysis of prevalence and validation

For validation purposes, the prevalence of conditions based on the final code lists was then cross-checked with prevalence estimates from relevant publications and the Quality and Outcomes Framework (a national system designed to remunerate general practices for

Table 2 Summary of code lists for selection criteria

Selection criterion (CPRD Aurum)	No. of codes	Appendix table
Diagnosis of AF (previous, current, or chronic)	43	S1
Existing use of an anticoagulant		
Oral anticoagulants	54	S2
Low molecular weight heparin	80	S3
Stroke	120	S4
Transient ischaemic attack	11	S5
Arterial thromboembolism	81	S6
Stroke risk factors		
Myocardial infarction	80	S7
Peripheral arterial disease	16	S8
Aortic disease	4	S9
Diabetes mellitus	455	S10
Antidiabetic drug/insulin	433	S11
Hypertension	85	S12
Antihypertensive drug	724	S13
Heart failure	54	S14
Loop diuretic therapy	59	S15
Prior major bleeding (intracranial bleed, or requiring hospitalization)	161	S16
High bleeding risk		
Gastrointestinal tract ulcer	234	S17
Brain injury	568	S18
Spinal injury	1	S19
Eye injury	11	S20
Estimated glomerular filtration rate <30 mL/min	6	S21
Azole-antimycotic treatment	20	S22
Current diagnosis of dementia	194	S23

providing good quality care to their patients). Results were discussed with the DaRe2THINK Data Coordination Team, and then suggested changes iteratively fed back to Phases 1–3 as appropriate. This process will be updated regularly through implementation so that phenotype definitions can be adjusted accordingly, for example if there is a change to coding systems. Cross-validation of endpoints will be conducted in a prospectively registered Study-Within-A-Trial.[27] In a subset of participants, we will compare outcome events obtained from coded primary and secondary EHR data with in-depth and granular information extracted from one of the UK's largest hospital trusts (University Hospitals Birmingham NHS Foundation Trust). This will involve extraction of text using machine learning methods from letters, clinical notes, imaging, and time-series data to determine accuracy and missingness.

Results and final code lists

Using the systematic approach discussed above, we searched a total of 1 176 611 codes, including 1 159 849 from primary care, and 16 762 from secondary care and national databases. This led to the creation of a total of 77 code lists across the different sections of the trial.

Participant selection criteria

The DaRe2THINK trial uses automated pre screening of the primary-care EHR to determine inclusion and exclusion according to the defined selection criteria. *Table 2* lists the main components of the selection criteria and associated CPRD Aurum code lists. A total of 3494 codes were included in these code lists using the systematic approach, including contribution from 14 published code lists.[5,29] Individual codes for each list can be found in Supplementary material online, *Appendix S1 Tables S1–S23*.

Table 3 Summary of code lists for outcome events

Outcome event	CPRD Aurum			HES		
	Number of codes reviewed	Final number of codes	Appendix table	Number of codes reviewed	Final number of codes	Appendix table
Cardiovascular mortality	–	–	–	423	353	S24
Ischaemic stroke	677	315	S25	41	10	S26
Transient ischaemic attack	62	44	S27	41	30	S28
Pulmonary and venous thromboembolism	1759	212	S29			
Arterial thromboembolism	1759	306	S30			
Thromboembolic event				144	67	S31
Myocardial infarction	215	150	S32	22	19	S33
Heart failure	430	100	S43	22	10	S44
Vascular dementia	252	58	S34	6	6	S35
Gastrointestinal bleeding	902	558	S36	42	25	S37
Other bleeding	4088	1642	S38	235	59	S39
Hospitalization event	3701	483	S40	–	–	–
Intracranial bleeding	1405	566	S41	36	31	S42

Table 4 Summary of code lists for baseline characteristics

Characteristic (CPRD Aurum)	Number of codes	Appendix table
Ethnicity	327	S45
Smoking status	104	S46
Measurements (height, weight, body mass index, blood pressure, and heart rate)	58	S47
Blood tests (creatinine, eGFR, haemoglobin, HbA1c, and lipids)	117	S48
AF	43	S1
Prior AF ablation	42	S49
Heart failure	100	S43
Hypertension	85	S12
Myocardial infarction	80	S32
Stroke	120	S4
Venous thromboembolism	212	S29
Arterial thromboembolism	306	S30
Peripheral artery disease and aortic plaque/atherosclerosis	20	S8 and S9
Diabetes mellitus	455	S10
Hyperthyroidism	73	S50
Hypothyroidism	90	S51
Chronic obstructive pulmonary disease	191	S52
Eye-related diseases and procedure	293	S53–S60
Gastrointestinal bleeding	558	S36
Intracranial bleeding	566	S41
Other bleeding	1642	S38
Hospitalization	483	S40
Antiplatelet therapy	79	S61
Diuretics	243	S15 and S62–S64
Calcium channel blockers	267	S65 and S66
ACE inhibitors	54	S67
Angiotensin II receptor antagonists	116	S68
Beta-blockers	115	S69
Alpha-blockers	58	S70
Aldosterone receptor antagonist	32	S71
Other antihypertensive drugs	76	S72
Antiarrhythmic drugs (Classes 1 and 3)	94	S73 and S74
Digoxin	14	S75
Sodium–glucose transport protein 2 inhibitors	34	S76
Non-steroidal anti-inflammatory drugs	86	S77

ACE, angiotensin-converting enzyme.

Outcome events and safety search

Outcome events in the DaRe2THINK trial are collated from primary-care, secondary-care, and national mortality data. Automated safety searches are also performed on a monthly basis in the primary-care EHR. *Table 3* provides a summary of the code lists used for outcome events, including 4434 codes from CPRD Aurum and 610 codes from HES and ONS. For safety events, 3081 CPRD Aurum codes were used to determine ischaemic stroke, intracranial bleeding, gastrointestinal bleeding, and other bleeding, with 483 codes to define associated hospitalization. Code lists were supplemented after review of 27 code lists from existing published literature,[30–43] with the final list of individual codes detailed in Supplementary material online, *Appendix S1, Tables S24–S44*.

Baseline characteristics

To remove the need for case report forms, baseline characteristics in the DaRe2THINK trial are extracted from the primary-care EHR. *Table 4* provides a summary of these code lists, with a total of 6802 codes from CPRD Aurum. Measurements and blood results are extracted as the latest value in the past 12 months, medications during the last 12 months, and health conditions throughout the participant's primary-care EHR. For heart failure, hypertension, and diabetes, diagnosis is classified in two ways: (i) use of one or more of the relevant codes identified in the EHR and (ii) a 'secure' diagnosis based on the code identified *in addition* to an associated medication in the last 12 months (loop diuretic therapy, antihypertensive medications, or oral antidiabetic/insulin, respectively). Codes are provided in Supplementary material online, *Appendix S1, Tables S1–S77*.

Discussion

The DaRe2 framework was designed to link together a series of innovations for a digital clinical trial, creating a patient-centred approach with high quality output, yet low burden on healthcare staff. The ability to pre screen from over 13 million NHS patient records allows for targeted and efficient recruitment, and simplifies enrolment of patients that are representative of the real-world population. Combining all health data from primary and secondary care sources nationally enables a 'remote' RCT, with no requirement for baseline or follow-up visits. These processes are supported by advances in digital technology, including remote e-consent and patient reported outcomes. The DaRe2THINK trial is using this sequence of innovations to test a hypothesis of critical importance to public health, with the aim of preventing the long-term consequences of AF, including cognitive decline and vascular dementia.

This study outlines our approach to defining phenotypes that support all aspects of the trial, from understanding the characteristics of recruited participants, through to determination of clinical endpoints. Unlike conventional trials with case report forms, DaRe2THINK extracts all relevant information from structured and coded healthcare data sources. Hence, it is critical for dissemination and transparency that all coding schema for this trial are pre published and available for evaluation (plus re-use) by other researchers and clinicians. In a published review of 450 EHR-based studies, only 19 (5.1%) were accompanied by a full set of clinical codes,[44] severely limiting the value of those studies, and their implementation to routine care. The emergence of EHR systems across the world provides an important opportunity for healthcare-embedded clinical research,[1] but only when this is accompanied by a clear pathway from data collection to interpretation.[2]

The broad scope of coded healthcare data has been a limiting factor in prior attempts at EHR-embedded clinical trials. As evidenced by the large number of codes screened for this trial (over a million),

a robust and systematic approach is required to ensure that patient events are not missed. The UK NHS is ideally suited to operationalize this sort of innovative trial, with a publicly funded health system that removes opportunity for personal financial gain from over or under-coding, while having systems in place to incentivize and monitor accurate coding. Structured healthcare data are used for quality and reimbursement purposes[4] and there is an ability to link across different healthcare sources.[45] Linkage is critical to understanding the full healthcare utilization of each participant,[3] in this case, combining historical and future primary-care, secondary-care, and death data. EHR-based trials may have an advantage in this regard compared with traditional trials, where investigators are relied on to document safety and outcome events. In the modern era of multiple heath providers for a multitude of different health conditions, sourcing safety and outcome events from national data have the potential to avoid missing endpoints, and potentially eliminate ascertainment bias in unblinded research. Although the value of trials based on routine EHR data is likely to vary depending on the disease being researched,[46] it is clear that the direction of travel is to enhance existing clinical trial infrastructure with digital innovations. In this regard, DaRe2THINK is the vanguard for cost-efficient RCTs embedded in clinical care. Previous trials have shown that although endpoints based on structured healthcare data may not replicate the traditional adjudication committee output, the estimated intervention effect is almost identical for adjudicated compared with EHR follow up.[47,48]

There remain substantial limitations to effective use of EHR data for clinical trials, including ongoing changes to coding systems. For example, ICD-10 includes 68 000 codes (a number of countries have their own editions, with ICD-11 launched this year), and SNOMED CT has over 300 000 clinical concepts. New codes can also be created within primary healthcare databases, highlighting the importance of iterative review and updates as exemplified by our coding framework. Universal limitations, such as missing data, variation in data quality, reliability of coding, and imprecision in the choice of code used, are common across real-world data sources. These and specific areas of bias such as lack of representativeness and loss to follow up are substantially minimized in DaRe2THINK due to the NHS being the de facto provider of almost all healthcare in the UK, and CPRD providing broad and representative coverage. In this setting, concordance between primary and secondary care health data is high. For example, in a random sample of 50 000 patients in CPRD Aurum, 94% of the 1260 patients with a code for myocardial infarction had corroborating evidence in the secondary care HES records.[3] Similar to our transparent approach to coding, the process of integrating primary and secondary healthcare sources in DaRe2THINK will be documented in an open-access, prospectively published Statistical Analysis Plan.

The DaRe2 approach was only possible with the support and guidance of our PPI team. We are indebted to the group for their constructive criticism throughout the design process, and their ongoing contribution to trial management to produce sustainable systems focused on the needs and wants of patients. Achieving a social licence for data-related clinical research is critical, and only achievable with high-level involvement by the public. In this case, we have used the PPI-POSITIVE approach,[8] within the context of the CODE-EHR best practice framework for using healthcare data in clinical research.[2]

Conclusions

The DaRe2THINK trial is using a series of digital innovations to reshape the deployment of an adaptive randomized clinical trial embedded in routine healthcare, with minimal burden for staff and patients. Integrating national healthcare data from primary and secondary sources, the system will provide evidence for a key public health concern, the prevention of cognitive decline and dementia in the rapidly growing number of patients with AF. Systematic and transparent methodology in the use of structured healthcare data, together with a patient and public mandate, are critical for evolution of these novel approaches in order to improve healthcare in our communities.

Author contributions

The coding framework was designed and implemented by X.W., O.T., K.O., K.N., D.S., and D.K. The manuscript was drafted by X.W., A.R.M. and D.K., with all other authors editing the manuscript for intellectual content. D.K. provided supervision and was responsible for the decision to submit the manuscript. This study is based in part on data from the Clinical Practice Research Datalink obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. The data are provided by patients and collected by the NHS as part of their care and support. The interpretation and conclusions contained in this study are those of the author/s alone.

Ethical approval and registration

The trial has received ethical/Health Research Authority approval (REC number: 21/NE/0021; IRAS project ID: 290420), regulatory approval (MHRA CTA 21761/0364) and is registered at Clinicaltrials.gov (NCT04700826), ISRCTN (21157803), and EudraCT (2020-005774-10).

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health* online.

Acknowledgements

In memory of Jaqueline Jones, member of the Patient and Public Involvement team. The authors acknowledge all members of the DaRe2THINK team and trial committees:

Trial Management Group: Susan Beatty (Clinical Practice Research Datalink, MHRA), Samir Mehta (Birmingham Clinical Trials Unit), Sophie Breeze (Clinical Practice Research Datalink, MHRA), Karen Lancaster (Clinical Practice Research Datalink, MHRA), and Stuart Fordyce (Clinical Practice Research Datalink, MHRA).

Data Coordination Team: Naomi Allen (University of Oxford, UK Biobank), Melanie Calvert (Centre for Patient Reported Outcomes Research and Birmingham Health Partners Centre for Regulatory Science and Innovation), Alastair Denniston (University Hospitals Birmingham NHS Foundation Trust and University of

Birmingham), and George Gkoutos (University of Birmingham and University Hospitals Birmingham NHS Foundation Trust).

Health Economics Team: Sahan Jayawardana (London School of Economics and Political Science) and Elias Mossialos (London School of Economics and Political Science).

Expert Advisory Group: Simon Ball (University Hospitals Birmingham NHS Foundation Trust and Health Data Research UK Midlands), Colin Baigent (University of Oxford), Peter Brocklehurst (Birmingham Clinical Trials Unit), Will Lester (University Hospitals Birmingham NHS Foundation Trust), Richard McManus (University of Oxford), Stefano Seri (Aston University), and Janet Valentine (Innovate UK, UK Research and Innovation).

Independent Trial Steering Committee: A. John Camm (St. George's University of London), Sandra Haynes MBE (Patient and Public Involvement team lead), Dame Julie Moore (University of Warwick), Amy Rogers (University of Dundee), and Mary Stanbury (Patient and Public Involvement).

Independent Data Monitoring Committee: Marcus Flather (University of East Anglia), Suzy Walker (General Practitioner, Fortrose Medical Practice, Highland), and Duolao Wang (Liverpool School of Tropical Medicine).

Funding

The DaRe2THINK trial is funded by the National Institute for Health and Care Research (NIHR130280). Medications in the trial are funded by the UK Department of Health and Social Care. The opinions expressed in this study are those of the authors and do not represent the NIHR or the UK Department of Health and Social Care.

Conflict of interest: All authors have completed the ICMJE uniform disclosure form (www.icmje.org/coi_disclosure.pdf) and declare:

X.W. and A.R.M. posts are funded by the National Institute for Health and Care Research (NIHR) (130280). O.T. post was funded by the EU/EFPIA Innovative Medicines Initiative (BigData@Heart 116074) and a grant from Amomed Pharma, outside the submitted work. K.O. has nothing to disclose. R.E.G. reports grants from NIHR, during the conduct of the study. P.M. has nothing to disclose. T.W. has nothing to disclose. S.H. reports a grants from NIHR (130280), during the conduct of the study. K.N. reports grants from NIHR, during the conduct of the study; and research grants from UKRI/MRC, Kennedy Trust for Rheumatology Research, Health Data Research UK, Wellcome Trust, European Regional Development Fund, Institute for Global Innovation, Boehringer Ingelheim, Action Against Macular Degeneration Charity, Midlands Neuroscience Teaching and Development Funds, South Asian Health Foundation, Vifor Pharma, College of Police, and CSL Behring, outside the submitted work; and consulting fees from Boehringer Ingelheim, Sanofi, CEGEDIM, MSD and holds a leadership/fiduciary role with NICST, a charity and OpenClinical, a Social Enterprise; outside the submitted work. D.S. reports grants from NIHR (130280), during the conduct of the study; and grants from Pfizer outside the submitted work. D.K. reports grants from NIHR (NIHR CDF-2015-08-074 RATE-AF; NIHR130280 DaRe2THINK; NIHR132974 D2T-NeuroVascular), the British Heart Foundation (PG/17/55/33087, AA/18/2/34218 and FS/CDRF/21/21032), the EU/EFPIA Innovative Medicines Initiative (BigData@Heart 116074), and the European Society of Cardiology supported by educational grants from Boehringer Ingelheim/BMS-Pfizer Alliance/Bayer/Daiichi Sankyo/Boston Scientific, the NIHR/University of Oxford Biomedical Research Centre and British Heart Foundation/University of Birmingham Accelerator Award (STEEER-AF NCT04396418). In addition, he has received a research grant to his institution from Pfizer; and research grants and advisory board fees from Bayer, Amomed and Protherics Medicines Development; all outside the submitted work.

Data availability

No additional data are available at this time.

References

1. Studer R, Sartini C, Suzart-Woischnik K, Agrawal R, Natani H, Gill SK, Wirta SB, Asselbergs FW, Dobson R, Denaxas S, Kotecha D. Identification and mapping real-world data sources for heart failure, acute coronary syndrome, and atrial fibrillation. *Cardiology* 2022;**147**:98–106.
2. Kotecha D, Asselbergs FW, European Society of Cardiology, BigData@Heart consortium, CODE-EHR international consensus group. CODE-EHR best practice framework for the use of structured electronic healthcare records in clinical research. *BMJ* 2022;**378**:e069048.
3. Persson R, Sponholtz T, Vasilakis-Scaramozza C, Hagberg KW, Williams T, Kotecha D, Myles P, Jick SS. Quality and completeness of myocardial infarction recording in Clinical Practice Research Datalink Aurum. *Clin Epidemiol* 2021;**13**:745–753.
4. Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, Faiz OD. Systematic review of discharge coding accuracy. *J Public Health (Oxf)* 2012;**34**:138–148.
5. Persson R, Vasilakis-Scaramozza C, Hagberg KW, Sponholtz T, Williams T, Myles P, Jick SS. CPRD Aurum database: assessment of data quality and completeness of three important comorbidities. *Pharmacoepidemiol Drug Saf* 2020;**29**:1456–1464.
6. Kotecha D, Shukla D, Clinical Practice Research Datalink. Preventing stroke, premature death and cognitive decline in a broader community of patients with atrial fibrillation using healthcare data for pragmatic research: a randomised controlled trial (DaRe2THINK). https://www.birmingham.ac.uk/dare2think (30 March 2022).
7. Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, Myles P. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol* 2019;**48**:1740–1740g.
8. Bunting KV, Stanbury M, Tica O, Kotecha D. Transforming clinical research by involving and empowering patients - the RATE-AF randomized trial. *Eur Heart J* 2021;**42**:2411–2414.
9. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, Castella M, Diener HC, Heidbuchel H, Hendriks J, Hindricks G, Manolis AS, Oldgren J, Popescu BA, Schotten U, Van Putte B, Vardas P. 2016 ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *Eur Heart J* 2016;**37**:2893–2962.
10. Kotecha D, Holmes J, Krum H, Altman DG, Manzano L, Cleland JG, Lip GY, Coats AJ, Andersson B, Kirchhof P, von Lueder TG, Wedel H, Rosano G, Shibata MC, Rigby A, Flather MD, Beta-Blockers in Heart Failure Collaborative Group. Efficacy of beta blockers in patients with heart failure plus atrial fibrillation: an individual-patient data meta-analysis. *Lancet* 2014;**384**:2235–2243.
11. Tica O, Khamboo W, Kotecha D. Breaking the cycle of HFpEF and AF. *Card Fail Rev* 2022. doi:10.15420/cfr.2022.03.
12. Kotecha D, Breithardt G, Camm AJ, Lip GYH, Schotten U, Ahlsson A, Arnar D, Atar D, Auricchio A, Bax J, Benussi S, Blomstrom-Lundqvist C, Borggrefe M, Boriani G, Brandes A, Calkins H, Casadei B, Castella M, Chua W, Crijns H, Dobrev D, Fabritz L, Feuring M, Freedman B, Gerth A, Goette A, Guasch E, Haase D, Hatem S, Haeusler KG, Heidbuchel H, Hendriks J, Hunter C, Kaab S, Kespohl S, Landmesser U, Lane DA, Lewalter T, Mont L, Nabauer M, Nielsen JC, Oeff M, Oldgren J, Oto A, Pison L, Potpara T, Ravens U, Richard-Lordereau I, Rienstra M, Savelieva I, Schnabel R, Sinner MF, Sommer P, Themistoclakis S, Van Gelder IC, Vardas PE, Verma A, Wakili R, Weber E, Werring D, Willems S, Ziegler A, Hindricks G, Kirchhof P. Integrating new approaches to atrial fibrillation management: the 6th AFNET/EHRA Consensus Conference. *Europace* 2018;**20**:395–407.
13. Conen D, Rodondi N, Muller A, Beer JH, Ammann P, Moschovitis G, Auricchio A, Hayoz D, Kobza R, Shah D, Novak J, Schlapfer J, Di Valentino M, Aeschbacher S, Blum S, Meyre P, Sticherling C, Bonati LH, Ehret G, Moutzouri E, Fischer U, Monsch AU, Stippich C, Wuerfel J, Sinnecker T, Coslovsky M, Schwenkglenks M, Kuhne M, Osswald S. Relationships of overt and silent brain lesions with cognitive function in patients with atrial fibrillation. *J Am Coll Cardiol* 2019;**73**:989–999.
14. Islam MM, Poly TN, Walther BA, Yang HC, Wu CC, Lin MC, Chien SC, Li YC. Association between atrial fibrillation and dementia: a meta-analysis. *Front Aging Neurosci* 2019;**11**:305.

15. Kirchhof P, Haeusler KG, Blank B, De Bono J, Callans D, Elvan A, Fetsch T, Van Gelder IC, Gentlesk P, Grimaldi M, Hansen J, Hindricks G, Al-Khalidi HR, Massaro T, Mont L, Nielsen JC, Nolker G, Piccini JP, De Potter T, Scherr D, Schotten U, Themistoclakis S, Todd D, Vijgen J, Di Biase L. Apixaban in patients at risk of stroke undergoing atrial fibrillation ablation. *Eur Heart J* 2018;**39**:2942–2955.
16. Lane DA, Skjoth F, Lip GYH, Larsen TB, Kotecha D. Temporal trends in incidence, prevalence, and mortality of atrial fibrillation in primary care. *J Am Heart Assoc* 2017; **6**:e005155.
17. European Medicines Agency Committee for Medicinal Products for Human Use. Assessment report: Direct oral anticoagulants (DOACs). EMA/194375/2020; Procedure number: EMEA/H/A-5(3)/1487. 2020.
18. Kotecha D, Pollack CV Jr, De Caterina R, Renda G, Kirchhof P. Direct oral anticoagulants halve thromboembolic events after cardioversion of AF compared with warfarin. *J Am Coll Cardiol* 2018;**72**:1984–1986
19. National Institute for Health and Care Excellence. *Atrial fibrillation: diagnosis and management.* NICE clinical guideline 196. 2021. www.nice.org.uk/guidance/ng196.
20. Wise J, de Barron AG, Splendiani A, Balali-Mood B, Vasant D, Little E, Mellino G, Harrow I, Smith I, Taubert J, van Bochove K, Romacker M, Walgemoed P, Jimenez RC, Winnenburg R, Plasterer T, Gupta V, Hedley V. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. *Drug Discov Today* 2019;**24**: 933–938.
21. NHS Digital. SNOMED CT. 2021. https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct (11 May 2022).
22. NHS Digital. Read codes. 2020. https://digital.nhs.uk/services/terminology-and-classifications/read-codes (11 May 2022).
23. NHS Business Services Authority. Dictionary of medicines and devices (dm + d). 2021. https://www.nhsbsa.nhs.uk/pharmacies-gp-practices-and-appliance-contractors/dictionary-medicines-and-devices-dmd (11 May 2022).
24. World Health Organization. International classification of disease version 10. 2019. https://icd.who.int/browse10/2019/en#/XXI (11 May 2022).
25. NHS. OPCS classification of interventions and procedures. 2022. https://www.datadictionary.nhs.uk/data_elements/opcs-4_code.html (11 May 2022).
26. Office for National Statistics. User guide to mortality statistics. 2021. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/userguidetomortalitystatisticsjuly2017 (11 May 2022).
27. Wang X, Gkoutos GV, Kotecha D. SWAT 158: validation of clinical events in the DaRe2THINK data-enabled clinical trial. *MRC Northern Ireland Hub for Trials Methodology Research: Study-Within-A-Trial (SWAT) Repository Store*. 2019. https://www.qub.ac.uk/sites/TheNorthernIrelandNetworkforTrialsMethodologyResearch/FileStore/Filetoupload,1218963,en.pdf (May 2022).
28. Gokhale KM, Chandan JS, Toulis K, Gkoutos G, Tino P, Nirantharakumar K. Data extraction for epidemiological research (DExtER): a novel tool for automated clinical epidemiology studies. *Eur J Epidemiol* 2021;**36**:165–178.
29. Jick SS, Hagberg KW, Persson R, Vasilakis-Scaramozza C, Williams T, Crellin E, Myles P. Quality and completeness of diagnoses recorded in the new CPRD Aurum database: evaluation of pulmonary embolism. *Pharmacoepidemiol Drug Saf* 2020;**29**: 1134–1140.
30. Wu J, Mamas MA, Mohamed MO, Kwok CS, Roebuck C, Humberstone B, Denwood T, Luescher T, de Belder MA, Deanfield JE, Gale CP. Place and causes of acute cardiovascular mortality during the COVID-19 pandemic. *Heart* 2021;**107**:113–119.
31. McCormick N, Lacaille D, Bhole V, Avina-Zubieta JA. Validity of heart failure diagnoses in administrative databases: a systematic review and meta-analysis. *PLoS One* 2014;**9**:e104519.
32. McCormick N, Bhole V, Lacaille D, Avina-Zubieta JA. Validity of diagnostic codes for acute stroke in administrative databases: a systematic review. *PLoS One* 2015;**10**: e0135834.
33. Lix LM, Sobhan S, St-Jean A, Daigle JM, Fisher A, Yu OHY, Dell'Aniello S, Hu N, Bugden SC, Shah BR, Ronksley PE, Alessi-Severini S, Douros A, Ernst P, Filion KB. Validity of an algorithm to identify cardiovascular deaths from administrative health records: a multi-database population-based cohort study. *BMC Health Serv Res* 2021; **21**:758.
34. Nedkoff L, Lopez D, Hung J, Knuiman M, Briffa TG, Murray K, Davis E, Aria S, Robinson K, Beilby J, Hobbs MST, Sanfilippo FM. Validation of ICD-10-AM coding for myocardial infarction subtype in hospitalisation data. *Heart Lung Circ* 2022;**31**: 849–858.
35. Ando T, Ooba N, Mochizuki M, Koide D, Kimura K, Lee SL, Setoguchi S, Kubota K. Positive predictive value of ICD-10 codes for acute myocardial infarction in Japan: a validation study at a single center. *BMC Health Serv Res* 2018;**18**:895.
36. Nagamine T, Gillette B, Pakhomov A, Kahoun J, Mayer H, Burghaus R, Lippert J, Saxena M. Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data. *Sci Rep* 2020;**10**:21340.
37. Bosco E, Hsueh L, McConeghy KW, Gravenstein S, Saade E. Major adverse cardiovascular event definitions used in observational analysis of administrative databases: a systematic review. *BMC Med Res Methodol* 2021;**21**:241.
38. Miao B, Hernandez AV, Alberts MJ, Mangiafico N, Roman YM, Coleman CI. Incidence and predictors of major adverse cardiovascular events in patients with established atherosclerotic disease or multiple risk factors. *J Am Heart Assoc* 2020;**9**:e014402.
39. Groenendyk JW, Rivera AS, Sinha A, Lloyd-Jones DM, Feinstein MJ. Changes in proportionate cardiovascular mortality in patients with chronic infectious and inflammatory conditions in the United States, 1999–2018. *Sci Rep* 2021;**11**:23985.
40. Carrero JJ, de Jager DJ, Verduijn M, Ravani P, De Meester J, Heaf JG, Finne P, Hoitsma AJ, Pascual J, Jarraya F, Reisaeter AV, Collart F, Dekker FW, Jager KJ. Cardiovascular and noncardiovascular mortality among men and women starting dialysis. *Clin J Am Soc Nephrol* 2011;**6**:1722–1730.
41. McDonald L, Sammon CJ, Samnaliev M, Ramagopalan S. Under-recording of hospital bleeding events in UK primary care: a linked Clinical Practice Research Datalink and Hospital Episode Statistics study. *Clin Epidemiol* 2018;**10**:1155–1168.
42. Mehta NK, Abrams LR, Myrskyla M. US life expectancy stalls due to cardiovascular disease, not drug deaths. *Proc Natl Acad Sci U S A* 2020;**117**:6998–7000.
43. Miniño A, Klein R. *Health Mortality from Major Cardiovascular Diseases*. Health E-Stats (2007). Hyattsville, Maryland: National Center for Health Statistics; 2010.
44. Springate DA, Kontopantelis E, Ashcroft DM, Olier I, Parisi R, Chamapiwa E, Reeves D. Clinicalcodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. *PLoS One* 2014;**9**:e99825.
45. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, Goldman S, Janmohamed S, Kreuzer J, Leenay M, Michel A, Ong S, Pell JP, Southworth MR, Stough WG, Thoenes M, Zannad F, Zalewski A. Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017;**106**:1–9.
46. McCord KA, Ewald H, Agarwal A, Glinz D, Aghlmandi S, Ioannidis JPA, Hemkens LG. Treatment effects in randomised trials using routinely collected data for outcome assessment versus traditional trials: meta-research study. *BMJ* 2021;**372**:n450.
47. Kjoller E, Hilden J, Winkel P, Galatius S, Frandsen NJ, Jensen GB, Fischer Hansen J, Kastrup J, Jespersen CM, Hildebrandt P, Kolmos HJ, Gluud C, Group CT. Agreement between public register and adjudication committee outcome in a cardiovascular randomized clinical trial. *Am Heart J* 2014;**168**:197–204. e191–194.
48. Bowman L, Mafham M, Wallendszus K, Stevens W, Buck G, Barton J, Murphy K, Aung T, Haynes R, Cox J, Murawska A, Young A, Lay M, Chen F, Sammons E, Waters E, Adler A, Bodansky J, Farmer A, McPherson R, Neil A, Simpson D, Peto R, Baigent C, Collins R, Parish S, Armitage J. Effects of aspirin for primary prevention in persons with diabetes mellitus. *N Engl J Med* 2018;**379**:1529–1539.