Paper:

Song, J., Elliot, E., Morris, A., Kerssens, J., Akbari, A., Ellwood-Thompson, S. & Lyons, R. (2018). A case study in distributed team science in research using electronic health records. *The International Journal of Population Data Science*

http://dx.doi.org/10.23889/ijpds.v3i3.442

International Journal of Population Data Science

Journal Website: www.ijpds.org

A case study in distributed team science in research using electronic health records

Song, J[1*], Elliot, E[2], Morris, AD[2], Kerssens, JJ[3], Akbari, A[1], Ellwood-Thompson, S[1], and Lyons, RA[1]

Submission History

Submitted:	30/11/2017
Accepted:	01/06/2018
Published:	21/09/2018

[1]Farr Institute, Swansea University Medical School, Swansea, UK
[2]Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK
[3]Public Health & Intelligence, NHS National Services Scotland, Edinburgh, UK

Abstract

Introduction
Due to various regulatory barriers, it is increasingly difficult to move pseudonymised routine health data across platforms and among jurisdictions. To tackle this challenge, we summarized five approaches considered to support a scientific research project focused on the risk of the new non-vitamin K Target Specific Oral Anticoagulants (TSOACs) and collaborated between the Farr institute in Wales and Scotland.

Approach
In Wales, routinely collected health records held in the Secure Anonymous Information Linkage (SAIL) Databank were used to identify the study cohort. In Scotland, data was extracted from national dataset resources administered by the eData Research & Innovation Service (eDRIS) and stored in the Scottish National Data Safe Haven. We adopted a federated data and multiple analysts approach, but arranged simultaneous accesses for Welsh and Scottish analysts to generate study cohorts separately by implementing the same algorithm. Our study cohort across two countries was boosted to 6,829 patients towards risk analysis. Source datasets and data types applied to generate cohorts were reviewed and compared by analysts based on both sites to ensure the consistency and harmonised output.

Discussion
This project used a fusion of two approaches among five considered. The approach we adopted is a simple, yet efficient and cost-effective method to ensure consistency in analysis and coherence with multiple governance systems. It has limitations and potentials of extending and scaling. It can also be considered as an initialisation of a developing infrastructure to support a distributed team science approach to research using Electronic Health Records (EHRs) across the UK and more widely.

Keywords
Team science, cross-jurisdictional data linkage, electronic health records

Introduction

There is an increasing trend of conducting health research by using data linkage of electronic records. Data linkage techniques boost deeper analyses on data merged from information contained in separate datasets regarding same individuals (1,2). Health data linkage frameworks are well-established in a number of countries/regions, i.e. in Australia (3), Scotland (4), England (5) and Wales (6,7,8). There is limited literature on the practicalities of using linked data from different centres, countries and resources for research. Many research projects require data from multiple jurisdictions to obtain sufficient power to answer scientific questions. These projects can benefit from larger sample sizes for greater statistical power, especially for small number of exposures, rare conditions or outcomes (9,10,11); ascertain complete patient pathways, care and outcome; accurate data for longitudinal studies; cross-jurisdictional comparisons and so on (12). However, due to various regulatory barriers, it is increasingly difficult to move pseudonymised routine data across platforms and among jurisdictions. Challenges arising from using data from multiple jurisdictions, i.e. legal issues, organisation capacity, financial cost, separation of roles and data ownership have been partially addressed in the US, Australia and other European countries (13,14,15,16). An infrastructure in Australia was proposed for cross-jurisdictional health data linkage research across states to improve the quality of population research data (17). It has been implemented in various scientific studies (18,19). The current environment, in the US, is characterized by budget and technical challenges, but investments

[*]Corresponding Author:
Email Address: jiao.song@swansea.ac.uk (J Song)

http://dx.doi.org/10.23889/ijpds.v3i3.442

in data infrastructure are arguably cost-effective (20).

In 2014, a UK research collaboration, the Farr Institute of Health Informatics Research, was established, comprising four centres distributed across the UK (North of England, Wales, London and Scotland) (21). Within the Farr Institute, we are motivated to make a step further towards the development of an infrastructure that allows for, and supports, cross-country research within the UK and across the EU using Electronic Health Records (EHRs). A recent scientific research project collaborated between centres in Wales and Scotland focused on the risk of the new non-vitamin K Target Specific Oral Anticoagulants (TSOAC) to a certain group of patients. This was part of an EU project, led from the Farr Institute in Scotland at Edinburgh and which the Farr Institute Wales at Swansea agreed to join to develop and demonstrate cross-UK data integration and analysis. This project is also in the process of including other European jurisdictions. Several recent randomised controlled trials have provided strong evidence supporting the safety and efficacy of TSOACs compared to the vitamin K antagonist warfarin for patients with atrial fibrillation (AF). TSOACs have a favourable risk–benefit profile, with significant reductions in stroke, intracranial haemorrhage (ICH), and mortality, and with similar major bleeding as for warfarin, but increased gastrointestinal bleeding (22). However, the cost-effectiveness of TSOACs is debated (23). TSOAC antidotes for reversal of bleeding are not yet available, and the safety and efficacy of TSOACs are unclear in patients who were not included in the randomised controlled trials but who clinicians feel may benefit. The safety issue of TSOACs in people who have had an ICH is uncertain and requires investigation. Due to the relative rarity of these circumstances, this safety issue needs to be addressed with data from more than one country.

We considered multiple approaches to achieve projects requiring multi-site analysis, such as the TSOAC study, and summarised their advantages and disadvantages, respectively. The summarization is presented in Table 1.

In this paper, we report our approach in support of the European scientific research project as a case study of cross centre data-intensive research.

Approach

The first priority of the TSOAC study was to generate Welsh and Scottish cohorts using the same algorithm and appropriate source data from both sites respectively within a limited period.

Data location and access

In Wales, routinely collected health records held in the Secure Anonymous Information Linkage (SAIL) Databank were used to identify the study cohort and follow up the patients in this cohort. The SAIL Databank is a safe haven for billions of records on over 5 million living and deceased people over the population of Wales, with a complete data linkage and analysis toolset (see (2,6,7,24)). SAIL has a governance procedure model with an extremely fast approval rate compared to similar facilities. This has been achieved through the development of the Information Governance Review Panel (IGRP), which implements NHS ethics guidance on the use of de-identified data for research. When an organisation agrees to share data, they may choose to delegate due diligence on governance and the use of their data to the IGRP. When a researcher requires that data, approval is given directly by the IGRP on behalf of the original data producers. Hence, requests for multiple datasets can be handled quickly by a single body with the delegated authority to make decisions. The IGRP consists of representatives from the British Medical Association (BMA), National Research Ethics Service (NRES), Public Health Wales NHS Trust, NHS Wales Informatics Service (NWIS), and members of the public from the Consumer Panel (1). An IGRP application with supportive information was approved for the TSOAC project. This approval provided analysts in Wales and Scotland access to de-identified data from the Patient Episode Database for Wales (PEDW), Welsh General Practice dataset (WLGP), and Annual District Death Extract (ADDE) also known as Office of National Statistics (ONS) mortality, held in the SAIL Databank.

Scotland does not have a single comprehensive national data warehouse. Instead, data, under the responsibility of different data controllers, are held at both regional and national levels and subsets (groups of variables) can be brought together when there are clear research questions that have public benefit. The Public Benefit and Privacy Panel for Health and Social Care (PBPP) is a governance structure of NHS Scotland that was established with delegated authority from NHS Scotland (NHSS) Chief Executive Officers and the Registrar General. The PBPP has a formal mandate to scrutinise any request to use NHSS-controlled data and the NHS Central Register data controlled by the Registrar General. The committee balances the benefits of undertaking research against the potential risks to individuals' privacy. The administrative process that supports decision-making is layered to ensure that decisions are made in a timely manner. A PBPP request, supported by evidence of prior Ethics Committee approval, was successfully granted for this collaborative TSOAC study. Analysts based in Wales then gained access to the TSOAC project in the Scottish National Data Safe Haven. Source data was extracted from national dataset resources administered by the eData Reseach & Innovation Service (eDRIS) (25).

Cohort generation

Based on the accesses granted and existing governance systems of each safe haven, we initially considered approaches 2 and 3, described in Table 1. Then the challenge was how to tackle the remaining disadvantages of both approaches, i.e., harmonization of analysis strategies between multiple analysts and long learning curves. In our experience of multi-site replication of research, it can be time consuming, as descriptions of the generation of variables are often incomplete and substantial amount of iterations are required. To shorten this step, we arranged for the Swansea and Edinburgh analysts to have simultaneous access to each other's data (fusion of approaches 2 and 3), which allowed for real-time viewing, creation of analytical codes and live discussion on how to tackle these practical challenges. The conveniences of how real-time communication on multiple screens worked effectively shortened the learning curves of both analysts.

To construct the study cohort, the first step was to identify patients who had experienced a non-traumatic intracranial haemorrhage categorised from hospital admissions information between 30/08/2013 and 30/06/2015. In Wales, the source dataset used was PEDW, and in Scotland, the General/Acute Inpatient and Day Case (SMR01) dataset. Analyses of the data in both centres used International Classification of Disease 10th version (ICD-10) diagnostic codes. Linkage to community prescription data identified those patients who were subsequently administered an anticoagulant during the 90 days after an index hospital discharge for non-traumatic intracranial haemorrhage. The datasets used to identify anticoagulant prescriptions were WLGP dataset in Wales and the Prescribing Information System (PIS) in Scotland. The named anticoagulants studied were: TSOACs (rivaroxaban, dabigatran and apixaban) and warfarin (included as reference). Subsequent mortality was identified through linkage to national mortality records, which provide date and underlying cause of death within the follow-up period (Welsh ADDE dataset and National Records of Scotland (NRS) death records). The primary outcome was the first subsequent hospital admission for a Serious Vascular Event (SVE), including ischemic stroke, systemic embolism, intracranial haemorrhage, or extracranial haemorrhage within 1 year (follow-up period) from the index discharge date. All available individual-level data in Scotland for the case study is held in the Scottish National Data Safe Haven, with the remaining study data (Welsh) held in the SAIL Databank. The cohort generation algorithm is summarised in Figure 1.

The Welsh and Scottish cohorts consisted of 2,676 and 4,153 patients respectively, as can be viewed in Table 2.

By applying the same cohort generation algorithm across two countries, our study cohort was boosted to 6,829 patients towards risk analysis.

Learning outcomes

Tables 3 and 4 present the differences in variables and their definitions between the two systems.

Adopting a fusion of approaches 2 and 3 enables real-time viewing, editing and communication. Source datasets and data types applied to generate Welsh and Scottish cohorts were reviewed and compared by analysts based on both Swansea and Edinburgh sites to ensure the consistency. Elements that needed to be handled differently in two safe havens were investigated and solutions identified. Based on these investigation results, an R script (see Appendix 1) was generated for this study to manipulate datasets in both the Welsh and Scottish safe havens to be able to produce a harmonised output suitable for combination to answer the required research questions given access and resources.

Discussion

After considering the options, this project took the approach of a fusion of approaches 2 and 3, as there were existing analysts at two centres, with each analyst accessing each distributed system simultaneously, harmonising variables, co-writing analytical scripts and combining the outputs. This was the quickest and easiest method to ensure that consistency was embedded in our analysis while working within the existing governance systems of each safe haven. Too much extra, non-project related resources, activities and agreements needed to be achieved to reach the same final outcomes with a centralised data (approach 1) as the governance, security, IT and approvals structures in each safe haven would have required wider approvals and changes to existing implementations. This would have added considerable extra work in Scotland as it already had an existing approved project at the start of the process. To adopt a linked federated data and analysis approach (approach 4), each site has to trust the central site with established security protocols and governance approval. There is no example of implementing this approach in the UK yet. Providing datasets are completely harmonised (not the case) to ensure the consistency of the analysis, a distributed query approach (approach 5) could have been undertaken and would probably have been acceptable to the governance managements under which the project was executed. The barrier to this approach is the lack of a generic distributed query engine as these types of technologies tend to be very bespoke and focused on specific research projects and their objectives.

Many research projects require data and rapid harmonisation of methods from more than one country or region to promote and enable research. Our view is that using remote access to data from distributed researchers, data visualisation and real-time co-written analytical scripts can significantly improve the efficiency of replication studies. The approach we adopted is a simple, yet very efficient and cost-effective method to ensure consistency in analysis and coherence with multiple governance systems. The algorithm developed from this study for manipulating and combining datasets in both Welsh and Scottish safe havens is limited to the relevant datasets and variables used for this study. However, it is easily extended and scalable to all available data, providing sufficient time and resources are made available. While this project is in the process of including other European jurisdictions to answer the specific scientific questions, our approach can also be considered as an initialisation of a developing infrastructure to support a distributed team science approach to research using EHRs across the UK and more widely.

Acknowledgements

We acknowledge the support from The Farr Institute of Health Informatics Research. The Farr Institute is supported by a 10-funder consortium: Arthritis Research UK, the British Heart Foundation, Cancer Research UK, the Economic and Social Research Council, the Engineering and Physical Sciences Research Council, the Medical Research Council, the National Institute of Health Research, the National Institute for Social Care and Health Research (Welsh Assembly Government), the Chief Scientist Office (Scottish Government Health Directorates), the Wellcome Trust, (MRC Grant Nos: CIPHER MR/K006525/1, Scotland MR/K007017/1. This work was partially supported as a Healthcare Innovation Activity funded by EIT Digital, a Knowledge and Innovation Community of the European Institute of Innovation and Technology.

This study makes use of anonymised data held in the Secure Anonymised Information Linkage (SAIL) Databank. We would like to acknowledge all the data providers who make anonymised data available for research.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

References

1. Green E, Ritchie F, Mytton J, Webber DJ, Deave T, Montgomery A, et al. Enabling data linkage to maximise the value of public health research data. [Online]. London; 2015 [cited 2017 November 18. Available from: http://eprints.uwe.ac.uk/25387.

2. Jones KH, Ford DV, Jones C, Dsilva R, Thompson S, Brooks C, et al. A case study of the Secure Anonymous Information Linkage (SAIL) Gateway: a privacy-protecting remote access system for health-related research and evaluation. Journal of biomedical informatics. 2014 Aug 31; 50: p. 196-204. https://doi.org/10.1016/j.jbi.2014.01.003

3. Kelman C, Bass A, Holman C. Research use of linked health data—a best practice protocol. Australian and New Zealand journal of public health. 2002 Jun 1; 26(3): p. 251-5. https://doi.org/10.1111/j.1467-842X.2002.tb00682.x

4. Scottish Government. A blueprint for health records research in Scotland. [Online].; 2012 [cited 2017 November 18. Available from: http://www.scot-ship.ac.uk/sites/default/files/Reports/SHIP_BLUEPRINT_DOCUMENT_final_100712.pdf.

5. Gill L. OX-LINK: the Oxford medical record linkage system. In Record Linkage Techniques; 1997; Washington: The National Academies Press. p. 491. https://doi.org/10.17226/6491

6. Ford DV, Jones KH, Verplancke JP, Lyons RA, John G, Brown G, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. BMC health services research. 2009 Dec 1; 9(1): p. 157. https://doi.org/10.1186/1472-6963-9-157

7. Lyons R, Jones K, John G, Brooks C, Verplancke J, Ford D, et al. The SAIL databank: linking multiple health and social care datasets. BMC medical informatics and decision making. 2009 Dec 1; 9(1): p. 3. https://doi.org/10.1186/1472-6947-9-3

8. Jones KH, McNerney CL, Ford DV. Involving consumers in the work of a data linkage research unit. International Journal of Consumer Studies. 2014 Jan 1; 38(1): p. 45-51. https://doi.org/10.1111/ijcs.12062

9. Cohen J. A power primer. Psychological bulletin. 1992 Jul; 112(1): p. 155-159. http://dx.doi.org/10.1037/0033-2909.112.1.155

10. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hilsdale. NJ: Lawrence Earlbaum Associates; 1988. https://doi.org/10.1016/C2013-0-10517-X

11. Everitt B, Skrondal A. The Cambridge dictionary of statistics Cambridge: Cambridge University Press; 2002.

12. Boyd JH, Ferrante AM. Cross-Jurisdictional Linkage - Enabling research at the national level. [Online].; 2016 [cited 2017 November 25. Available from: http://www.med.monash.edu.au/assets/docs/creps/2016/2016datalinkageaug-annaferrante.pdf.

13. Hogan R, Bullard C, Stier D, Penn M, Wall T, Cleland J, et al. Assessing cross-sectoral and cross-jurisdictional coordination for public health emergency legal preparedness. The Journal of Law, Medicine & Ethics. 2008 Mar 1; 36(s1): p. 36-52. https://doi.org/10.1111/j.1748-720X.2008.00258.x

14. Hyde J, Shortell S. The structure and organization of local and state public health agencies in the US: a systematic review. American journal of preventive medicine. 2012 May 31; 42(5): p. S29-41. https://doi.org/10.1016/j.amepre.2012.01.021

15. Andrew NE, Sundararajan V, Thrift AG, Kilkenny MF, Katzenellenbogen J, Flack F, et al. Addressing the challenges of cross-jurisdictional data linkage between a national clinical quality registry and government-held health data. Australian and New Zealand journal of public health. 2016 Oct 1; 40(5): p. 436-42. https://doi.org/10.1111/1753-6405.12576

16. Tavares J, Oliveira T. Electronic Health Record Portal Adoption: a cross country analysis. BMC medical informatics and decision making. 2017 Jul 5; 17(1): p. 97. https://doi.org/10.1186/s12911-017-0482-9

17. Boyd JH, Ferrante AM, O'Keefe CM, Bass AJ, Randall SM, Semmens JB. Data linkage infrastructure for cross-jurisdictional health-related research in Australia. BMC health services research. 2012 Dec 1; 12(1): p. 480. https://doi.org/10.1186/1472-6963-12-480

18. Tran DT, Havard A, Jorm LR. Data cleaning and management protocols for linked perinatal research data: a good practice example from the Smoking MUMS (Maternal Use of Medications and Safety) Study. BMC medical research methodology. 2017 Dec 11; 17(1): p. 97. https://doi.org/10.1186/s12874-017-0385-6

19. Spilsbury K, Rosman D, Alan J, Ferrante AM, Boyd JH, Semmens JB. Improving the Estimation of Risk-Adjusted Grouped Hospital Standardized Mortality Ratios Using Cross-Jurisdictional Linked Administrative Data: A Retrospective Cohort Study. Frontiers in public health. 2017;: p. 5. https://doi.org/10.3389/fpubh.2017.00013

20. Bradley C, Penberthy L, Devers K, Holden D. Health services research and data linkages: issues, methods, and directions for the future. Health services research. 2010 Oct; 45(5p2): p. 1468-1488. https://doi.org/10.1111/j.1475-6773.2010.01142.x

21. Farr Institute. The Farr Institute of Health Informatics Research. [Online]. [cited 2017 November 19. Available from: http://www.farrinstitute.org/about/who-we-are.

22. Ruff CT, Giugliano RP, Braunwald E, Hoffman EB, Deenadayalu N, Ezekowitz MD, et al. Comparison of the efficacy and safety of new oral anticoagulants with warfarin in patients with atrial fibrillation: a meta-analysis of randomised trials. The Lancet. 2014 Mar 21; 383(9921): p. 955-62. https://doi.org/10.1016/S0140-6736(13)62343-0

23. Coyle D, Coyle K, Cameron C, Lee K, Kelly S, Steiner S, et al. Cost-effectiveness of new oral anticoagulants compared with warfarin in preventing stroke and other cardiovascular events in patients with atrial fibrillation. Value in health. 2013 Jun 30; 16(4): p. 498-506. https://doi.org/10.1016/j.jval.2013.01.009

24. SAIL Databank. Secure Anonymised Information Linkage (SAIL) Databank. [Online]. [cited 2017 November 17. Available from: https://saildatabank.com/.

25. Information Services Division Scotland, NHS National Services. eDRIS Products and services. [Online].; 2010 [cited 2017 November 6. Available from: http://www.isdscotland.org/Products-and-Services/EDRIS/.

Abbreviations

ADDE	Annual District Death Extract
AF	Atrial Fibrillation
BMA	British Medical Association
EHR	Electronic Health Record
ICD	International Classification of Disease
ICH	IntraCranial Haemorrhage
IGRP	Information Governance Review Panel
NRES	National Research Ethics Service
NRS	National Records of Scotland
NWIS	NHS Wales Informatics Service
ONS	Office for National Statistics
PBPP	Public Benefit and Privacy Panel for Health and Social Care
PEDW	Patient Episode Database for Wales
PIS	Prescribing Information System
SAIL	Secure Anonymous Information Linkage
SMR01	General/Acute Inpatient and Day Case
SVE	Serious Vascular Event
WLGP	Welsh Longitudinal General Practice dataset

Table 1: Summarization of considered approaches for a 3-centre analysis

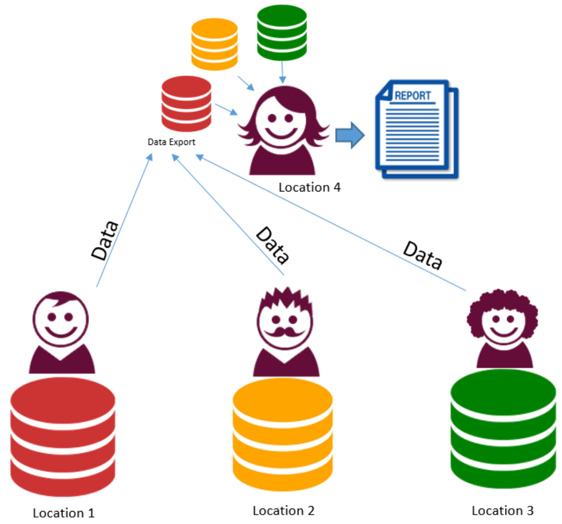
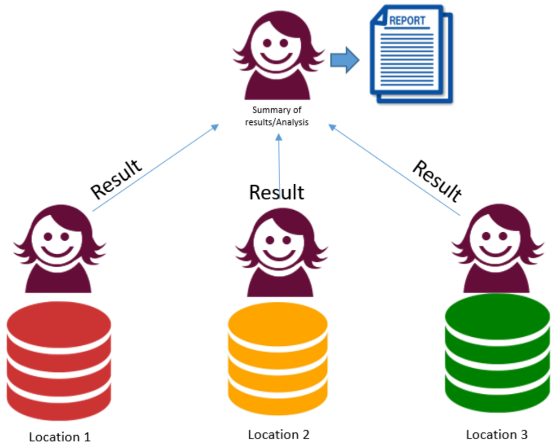
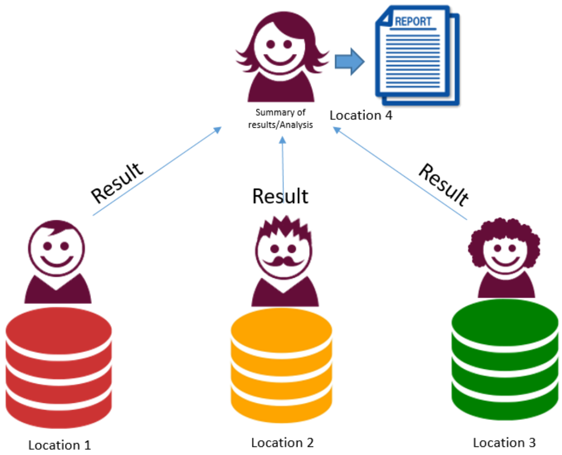
Approach	Advantages/Disadvantages
1. Centralised data and analysis. Data moved from 3 centres – 1 analyst (centralised data model)  REPORT Data Export Location 4 Data Data Data Location 1 Location 2 Location 3	All data submitted from each site to a central site. Advantages • Analysis is easier because of being fully in control of a single researcher having all the data available. Disadvantages • Each site must "trust" the central site and must seek governance approval from each site and perhaps put in place a legal contract. • Restrictions about data sovereignty may prevent this approach.
2. Federated data and single analyst. Data at 3 centres – 1 analyst accessing each platform then combining results  REPORT Summary of results/Analysis Result Result Result Location 1 Location 2 Location 3	Same researcher accesses each system separately and combines outputs. Advantages • Same researcher so same approach. Disadvantages • Access to separate systems and learning curves. • Separate access contracts and conditions of use. • If outputs need combined individual level analysis, then not workable.
3. Federated data and multiple analysts. Data at 3 centres – 3 separate analyses, combine results  REPORT Summary of results/Analysis Location 4 Result Result Result Location 1 Location 2 Location 3	All analysis done separately by host site with outputs collated. Advantages • Can be done quickly as each site knows their own system. Disadvantages • Consistency can be hard to achieve so more validation and process documentation required. • If outputs need combined individual level analysis, then not workable. • Dependant on resources available at each local site.

Table 1. Cont. Summarization of considered approaches for a 3-centre analysis

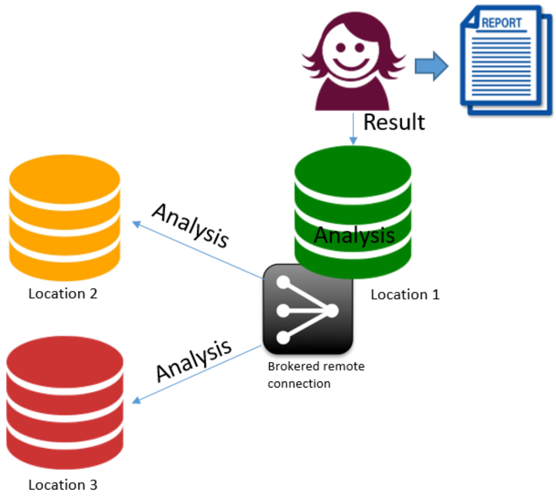
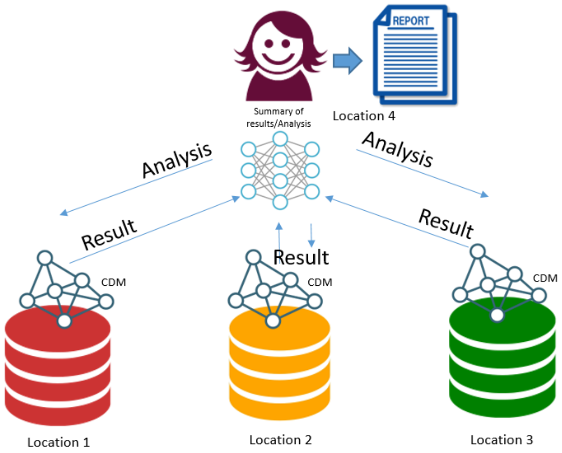
Approach	Advantages/Disadvantages
4. Linked federated data and analysis. Data at 3 centres – 1 analyst (remote real time access model)  REPORT Result Analysis Analysis Analysis Location 2 Location 1 Brokered remote connection Location 3	The sites have established inter connections. From a site the researcher can access all the required data. Advantages • Analysis is easier because of being fully in control of the researcher having all the data available. Disadvantages • Each site must "trust" the central site and must seek governance approval from each site and perhaps put in place a legal contract. • Restrictions about data sovereignty may prevent this approach.
5. Federated data and distributed analysis. Data at 3 centres – 1 analyst directing federated queries  REPORT Summary of results/Analysis Location 4 Analysis Analysis Result Result Result CDM CDM CDM Location 1 Location 2 Location 3	Using a distributed query system – issue same query to all sites. Advantages • Same analysis performed in each site. • No data moving so could be good for cross country restrictions. Disadvantages • Common data model required. • Data needs to be harmonised. • More complex from an IT and governance perspective.

Table 2: Study cohorts

	Welsh cohort	Scottish cohort	Total
Male	1,347	1,938	3,285
Female	1,329	2,215	3,544
Total	2,676	4,153	6,829

Figure 1: Cohort generation algorithm

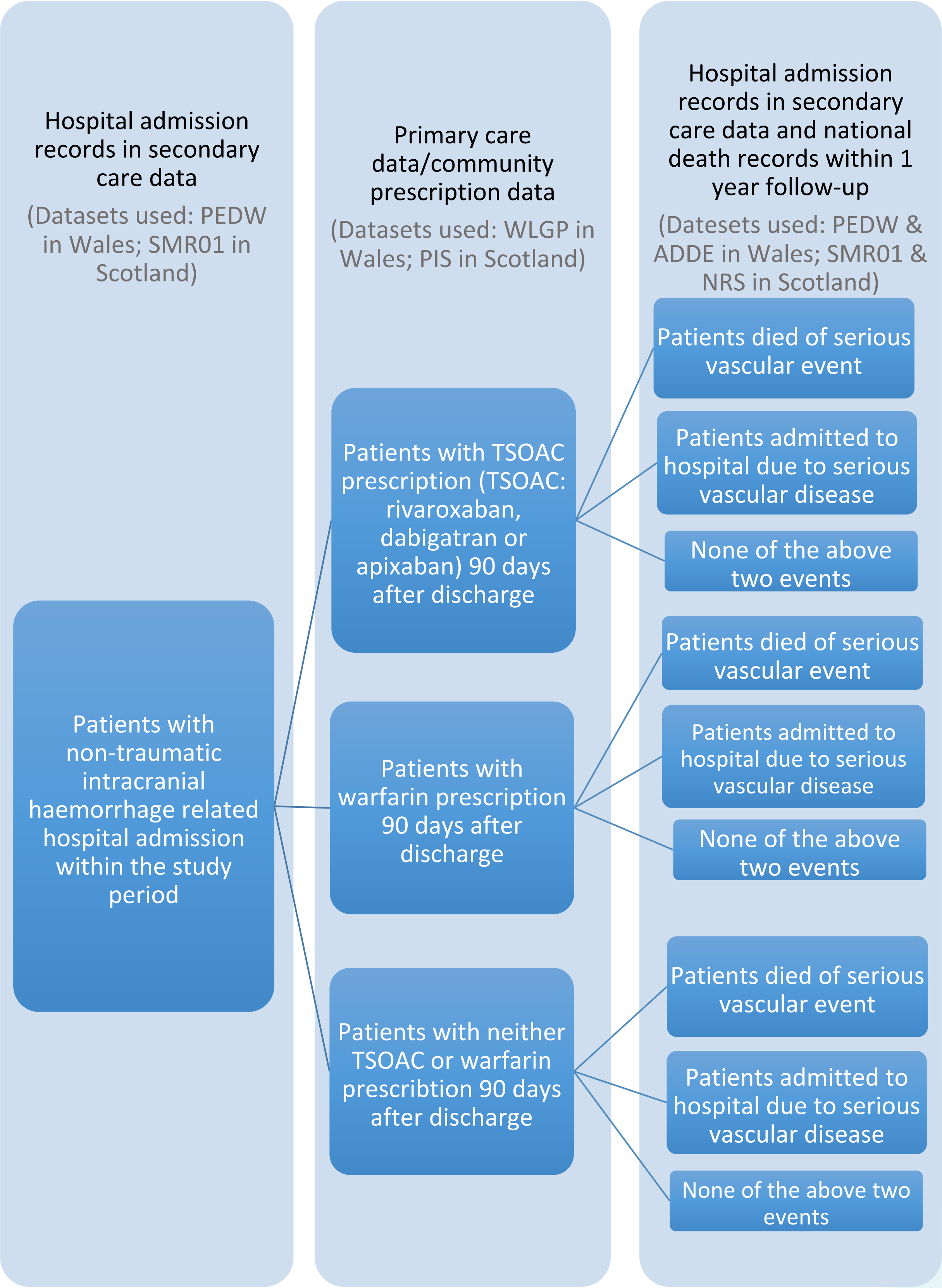

Hospital admission records in secondary care data
(Datasets used: PEDW in Wales; SMR01 in Scotland)
Primary care data/community prescription data
(Datasets used: WLGP in Wales; PIS in Scotland)
Hospital admission records in secondary care data and national death records within 1 year follow-up
(Datesets used: PEDW & ADDE in Wales; SMR01 & NRS in Scotland)
Patients with non-traumatic intracranial haemorrhage related hospital admission within the study period
Patients with TSOAC prescription (TSOAC: rivaroxaban, dabigatran or apixaban) 90 days after discharge
Patients with warfarin prescription 90 days after discharge
Patients with neither TSOAC or warfarin prescribtion 90 days after discharge
Patients died of serious vascular event
Patients admitted to hospital due to serious vascular disease
None of the above two events
Patients died of serious vascular event
Patients admitted to hospital due to serious vascular disease
None of the above two events
Patients died of serious vascular event
Patients admitted to hospital due to serious vascular disease
None of the above two events

Table 3: Variables comparison between Welsh and Scottish data

	Welsh cohort		Scottish cohort	
	Variable name	Source data	Variable name	Source data
Patient identity & linkage field	ALF_E	PEDW	UPI_NUMBER	SMR01
Admission date	ADMIS_DT	PEDW	ADMISSION_DATE	SMR01
Admission methods	ADMIS_MTHD_CD	PEDW	ADMISSION_TYPE	SMR01
Discharge types	DISCH_MTHD_CD	PEDW	DISCHARGE_TYPE	SMR01
Drugs prescription	EVENT_CD	WLGP	BNFItemcode	PIS
Date of prescription	EVENT_DT	WLGP	PRES_DATE	PIS
Date of birth	WOB	ADDE	DATE_OF_BIRTH	NRS
Gender	GNDR_CD	ADDE	SEX	NRS
Deprivation quintile	WIMD2011_5TH	PEDW	SIMD_QUINTILE	SMR01
Primary cause of death	DEATHCAUSE_DIAG_UNDERLYING_CD	ADDE	CAUSE_OF_DEATH_CODE	NRS
Date of death	DOD	ADDE	DATE_OF_DEATH	NRS

Table 4: Different data definitions in Wales and Scotland

		Welsh data	Scottish data
Gender	0	N/A	Not known (i.e. indeterminate sex, includes intersex)
	1	Male	Male
	2	Female	Female
	8	Not specified	N/A
	9	N/A	Not specified (includes not stated by patient, or not recorded)
Date	Date format	YYYY-MM-DD	DDMMYY
Drug	Drug information	EVENT_CD: READ codes, e.g. bs74.	British National Formulary Drug Codes (BNF). e.g. BNF item code 0601011A0BBADAC

Appendix 1

```
library(RODBC);
        ##EXTRACT PATIENTS ADMITTED TO HOSPITAL WITH ICH IN WALES
        ##ALF_STS_CD IN ('1','4','39') COVER ALF_Es WITH GOOD MATCHING
        ##SCORES
                ICHWales<- sqlQuery(sql,"SELECT ALF_E, PERSON_SPELL_ADMIS_DT,
                                PERSON_SPELL_DISCH_DT, PERSON_SPELL_NUM_E FROM (
                                SELECT B.ALF_E, A.PERSON_SPELL_ADMIS_DT,
                                A.PERSON_SPELL_DISCH_DT, A.PERSON_SPELL_NUM_E,
                                ROW_NUMBER () OVER (PARTITION BY B.ALF_E ORDER BY
                                A.PERSON_SPELL_ADMIS_DT) AS SEQ FROM (
                                (SELECT MIN(ADMIS_DT) AS PERSON_SPELL_ADMIS_DT,
                                MAX(DISCH_DT) AS PERSON_SPELL_DISCH_DT,
                                PERSON_SPELL_NUM_E
                                FROM
                                (
                                SELECT DISTINCT ALF_E, ADMIS_DT, DISCH_DT,
                                PERSON_SPELL_NUM_E
                                FROM PEDW
                                WHERE EPI_NUM=1
                                AND (DIAG_CD_123='I60'
                                OR DIAG_CD_123='I61'
                                OR DIAG_CD_123='I62')
                                AND
                                ALF_STS_CD IN ('1','4','39')
                                AND ALF_E IS NOT NULL
                                AND ADMIS_DT<='YYYY-MM-DD'
                                )
                                GROUP BY PERSON_SPELL_NUM_E )A
                                LEFT JOIN (
                                SELECT DISTINCT ALF_E, PERSON_SPELL_NUM_E FROM
                                PEDW
                                WHERE EPI_NUM=1
                                AND (DIAG_CD_123='I60'
                                OR DIAG_CD_123='I61'
                                OR DIAG_CD_123='I62')
                                AND
                                ALF_STS_CD IN ('1','4','39')
                                AND ALF_E IS NOT NULL
                                AND ADMIS_DT<='YYYY-MM-DD'
                                ) B
                                ON A.PERSON_SPELL_NUM_E=B.PERSON_SPELL_NUM_E
                                )
                                )
                                WHERE SEQ=1");


        ##LOOK FOR DEATH RECORDS FOR ANY OF THE PATIENTS WITH ICH RELATED
        ##HOSPITAL ADMISSION IN WALES
        ##THIS TABLE INCLUDES INDIVIDUAL INFORMATION OF PRIMARY AND
        ##SECODARY CAUSES OF DEATH
        DeathWales<- sqlQuery(sql,"SELECT * FROM ADDE WHERE ALF_E IN
                                (SELECT ALF_E FROM ICHpatients)");
```

```
##LOOK FOR ADMISSION RECORDS FOR ANY OF ICH PATIENTS RE-ADMITTED TO
##HOSPITAL WITH SERIOUS VASCULAR DISEASE
ReadmissionWales <-  sqlQuery(sql,"SELECT * FROM(
                        SELECT A.*, ROW_NUMBER () OVER
                        (PARTITION BY
                        A.ALF_E ORDER BY A.ADMIS_DT) AS SEQ  FROM(
                        SELECT * FROM PEDW
                        WHERE EPI_NUM=1
                        AND (substr(DIAG_CD_1234, 1, 3) IN
                        (SELECT ICD10
                        FROM RelaventICDcodes)
                        OR DIAG_CD_1234 IN (SELECT ICD10 FROM
                        RelaventICDcodes))
                        ) A
                        INNER JOIN ICHPATIENTSLIST B
                        ON A.ALF_E=B.ALF_E
                        AND B.PERSON_SPELL_DISCH_DT < A.ADMIS_DT
                        )
                        WHERE SEQ=1");

##ANY OF THESE PATIENTS HAVE BEEN PRISCRIDED TSOACS OR WARFARIN
##ONLY APIXABAN IS ILLUSTRATED HERE, SAME ALGOTHRIM CAN BE APPLIED TO
##OTHER DRUG PRESCRIBTIONS
APIXABANWales<- sqlQuery(sql,"SELECT A.ALF_E, A.PRAC_CD_E,
                        A.LOCAL_NUM_C
                        , A.EVENT_CD, A.EVENT_DT, A.EVENT_YR,
                        A.SOURCE_EXTRACT, B.CD_DESCRIPTION
                        FROM
                        (SELECT * FROM WLGP
                        WHERE EVENT_DT>='YYYY-MM-DD'
                        )A
                        INNER JOIN READCODESFORAPIXABANPRESCRIBTION B
                        ON A.EVENT_CD=B.READ_CD");


##MERGE TO PRODUCE WELSH STUDY COHORT
##AGAIM ONLY APIXABAN IS ILLUSTRATED HERE
ICHDeathWales <- merge(ICHWales, DeathWales, by="ALF_E");
ICHDeathReadmWales<- merge(ICHDeathWales, ReadmissionWales,
        by="ALF_E");
ICHDeathReadmAPIWales<- merge(ICHDeathReadmWales, APIXABANWales,
        by="ALF_E");
```

```
##EVENT WAS DEFINED AS DEATH OR READMISSION DUE TO SERIOUS VASCULAR
##EVENTS
##ALL TSOACS AND WARFRIN HAVE BEEN GROUPED TO ONE COLUMN CALLED
##DRUG WITH FLAGS
TSOACWales <- ICHDeathReadmAPIDABRIVWARWales#WELSH COHORT MEATATABLE
        [,c("ALF_E"
        ,"AGE"
        ,"AGE_GROUP"
        ,"GENDER"
        ,"WIMD2011_5TH"
        ."ADMIS_MTHD_CD"
        ,"ENTRY_DT" # ADMISSION DATE
        ,"DOD" #DATE OF DEATH
        ,’’DEATH_PRI_OUTCOME’’#FLAG WHETHER DEATH WAS DUE TO
        #SERIOUS VASCULAR EVENTS
        ,"READMIS_PRI"#FLAG WHETHER READMISSION WAS DUE TO
        #SERIOUS VASCULAR EVENTS
        ,"EVENT_DT"
        ,"EXPOSURE" #FLAG WHETHER TSOAC OR WARFARIN
        # PRESCRIBTION HAD BEEN GIVEN
        ,"EXPOSURE_DT"#DATE OF TSOAC OR WARFARIN
        #PRESCRIBTION
        ,"DRUG" #TSOAC, WARFARIN OR NEITHER
        )];

##CHANGE COLUMN NAMES
colnames(TSOACWales) <- c(
        "IDENTIFIER"
        ,"AGE"
        ,"AGE_GROUP"
        ,"GENDER"
        ,"DEPRIVATION"
        ."ADMISSION_METHOD"
        ,"ENTRY_DATE"
        ,"DEATH_DATE"
        ,"DEATH_PRIMARY_OUTCOME"
        ,"READMISSION_PRIMARY_OUTCOME"
        ,"EVENT_DATE"
        ,"EXPOSURE"
        ,"EXPOSURE_DATE"
        ,"DRUG"
        );

ICHScotland <- sqlQuery(sql,"SELECT TO.UPI, TO.ID_ISD, TO.INDEX_DATE,
        TO.INDEX_CIS,
        T1.ADMISSION_DATE, T1.ADMISSION_REASON,
        T1.ADMISSION_TYPE,
        T1.AGE_IN_YEARS, T1.DISCHARGE_TYPE,
        T1.DISCHARGE_DATE,
        T1.DR_POSTCODE, MANAGEMENT_OF_PATIENT,
        T1.CIS_MARKER,
        T1.MAIN_CONDITION, T1.OTHER_CONDITION_1,
        T1.OTHER_CONDITION_2, T1.OTHER_CONDITION_3,
        T1.OTHER_CONDITION_4, T1.OTHER_CONDITION_5,
        T1.MAIN_OPERATION, T1.OTHER_OPERATION_1,
        T1.OTHER_OPERATION_2, T1.OTHER_OPERATION_3,
        T1.SEX,
        T1.UPI_NUMBER FROM { UPIfile TO LEFT OUTER
        JOIN SMR01 T1
        ON TO.UPI = T1.UPI_NUMBER}
        ");
```

```
DeathScotland<- sqlQuery(sql,"SELECT T0.UPI, T0.ID_ISD,
        T0.INDEX_DATE, T1.AGE,
        T1.CAUSE_OF_DEATH_CODE_0,
        T1.CAUSE_OF_DEATH_CODE_1,
        T1.CAUSE_OF_DEATH_CODE_2,
        T1.CAUSE_OF_DEATH_CODE_3,
        T1.CAUSE_OF_DEATH_CODE_4,
        T1.CAUSE_OF_DEATH_CODE_5,
        T1.CAUSE_OF_DEATH_CODE_6,
        T1.CAUSE_OF_DEATH_CODE_7,
        T1.CAUSE_OF_DEATH_CODE_8,
        T1.CAUSE_OF_DEATH_CODE_9,
        T1.DATE_OF_DEATH, T1.SEX, T1.UPI_NUMBER
        FROM { UPIfile
        T0 LEFT OUTER JOIN
        NRS T1 ON T0.UPI = T1.UPI_NUMBER }
        ORDER BY T0.UPI ASC’");

DrugScotland<- sqlQuery(sql," SELECT PatUPIC
        PatPostcodeC A21
        PatCareHomeResidencyFlag A1
        PrescDate A19
        DispDate A19
        PIBNFItemDescription A23
        PIItemStrengthUOM A14
        PIDailyDose F1.0
        PIDailyDoseUOM F1.0
        PIApprovedName A15
        PIDailyDoseConversion F3.1
        PaidQuantity F3.0
        NumberofPaidItems F1.0.
        FROM PIS
        WHERE DATE #DATE HAS TO BE
        #WITHIN 90 DAYS
        #AFTER DISCHARE IN THIS STUDY
        AND UPI #PATIENTS WITH ICD
        # ADMISSION
        );

##MERGE TOGETHER TO OBATIN SCOTTISH SOTUDY COHORT
TSOACScotland <- ICHDeathReadmAPIDABRIVWARScotland #SCOTTISH COHORT
        [,c(
        "UPI_NUMBER"
        ,"AGE_IN_YEARS"
        ,"AGE_GROUP"
        ,"SEX"
        ,"SIMD"
        ,"ADMISSION_TYPE"
        ,"ENTRY_DATE"
        ,"DEATH_DATE"
        ,"CAUSE_OF_DEATH_CODE"
        ,"ADMISSION_DATE"
        ,"EVENT_DATE"
        ,"EXPOSURE"
        ,"EXPOSURE_DATE"
        ,"DRUG"
        )];
```

```
colnames(TSOACScotland) <- c(
        "IDENTIFIER"
        ,"AGE"
        ,"AGE_GROUP"
        ,"GENDER"
        ,"DEPRIVATION"
        ,"ADMISSION_METHOD"
        ,"ENTRY_DATE"
        ,"DEATH_DATE"
        ,"DEATH_PRIMARY_OUTCOME"
        ,"READMISSION_PRIMARY_OUTCOME"
        ,"EVENT_DATE"
        ,"EXPOSURE"
        ,"EXPOSURE_DATE"
        ,"DRUG"
        );

##IF TWO COHORTS CAN BE COMBINED, A FEW ADJUSTMENTS ARE NEEDED BEFORE
##UPDATE GENDER CODE. IN SCOTLAND FROM 0 AND 9 TO 8
TSOACScotland$GENDER[TSOACScotland$GENDER==0|TSOACScotland$GENDER==9]
        <- 8;

##UPDATE DATE IN SCOTLAND FROM DD/MM/YY TO DD-MM-YYYY
TSOACScotland$ENTRY_DATE <- as.Date(TSOACScotland$ENTRY_DATE, format=
        "%d%m%y");

TSOACScotland$DEATH_DATE <- as.Date(TSOACScotland$DEATH_DATE, format
        = "%d%m%y");

TSOACScotland$EVENT_DATE <- as.Date(TSOACScotland$EVENT_DATE, format
        = "%d%m%y");

TSOACScotland$EXPOSURE_DATE <- as.Date(TSOACScotland$EXPOSURE_DATE,
        format = "%d%m%y");

##ADD A FLAG INDICATE WHERE THE PATIENT CAME FROM
        TSOACScotland$LOCATION <- S
        TSOACWales$LOCATION <- W

        ##UNION WALESH AND SCOTTISH DATASETS
        TSOAC <- rbind(TSOACWales,TSOACScotland);
        ##GIVE STUDY_ID
        TSOAC$STUDY_ID <- seq.int(nrow(TSOAC));
        ##DROP THE COLUMN IDENTIFIER WHICH INCLUDES ALF_E AND UPI_NUMBER
        drop <- c("IDENTIFIER")
        TSOAC[ , !(names(TSOAC) %in% drop)]
```