CPRD OLD and Linked ONS Mortality Records: Reconciling uidelines

Antonella Delmestri[a,1], Daniel Prieto-Alhambra[a]

[a]*Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, U.K.*

[1]Corresponding author at: BOTNAR Research Centre, Nuffield Orthopaedic Centre, Windmill Road, OX3 7LD, Oxford, U.K.
E-mail address: antonella.delmestri@ndorms.ox.ac.uk

Abstract

Background

The Clinical Practice Research Datalink (CPRD) GOLD is an extremely influential U.K. primary care dataset for epidemiological research having a number of published papers based on its data much bigger than any other U.K. primary care dataset. The Office for National Statistics (ONS) death data for England can be linked to GOLD at the patient level and are considered the gold standard on mortality. GOLD, which also holds death data, has been recently assessed against ONS linked dataset and the accuracy of its dates of death has been deemed sufficient for the majority of observational studies. However, there is a lack of guidance on how to manage the challenges existing when ONS mortality and GOLD datasets are linked, including linkage coverage period, linkage correctness likelihood, linkage regional limitations and data discrepancy.

Objectives

Provide reconciling guidelines on how to make maximum and at the same time trustworthy use of mortality information coming from both GOLD and ONS linked datasets with the aim of improving the quality, reproducibility, transparency and comparison of clinical research.

Method and results

We have developed recommendations on how to manage mortality data coming from both GOLD and linked ONS, taking into account linkage coverage period, linkage correctness likelihood, linkage regional limitations and data discrepancies between these two datasets. We have also implemented these guidelines in an SQL algorithm for researchers to use.

Conclusion

We have provided detailed guidelines on the reconciliation of mortality data between GOLD and ONS linked death datasets, taking into account both their strengths and limitations. The

consistent application of these guidelines made practical by an SQL algorithm, has the potential to improve clinical research quality, reproducibility, transparency and comparison.

Keywords: CPRD GOLD, ONS, death, mortality, data linkage

1. Introduction

The Clinical Practice Research Datalink (CPRD) GOLD dataset [1] is the most influential U.K. primary care data source of anonymised medical records and it is widely used in epidemiological studies. The number of published papers based on its data is impressive [2] and much bigger than the sum of the papers based on all the other U.K. primary care datasets, despite representing at the time of writing only 4.26% of the current population and 3.79% of general practices [3]. Mortality information, and in particular date of death, is frequently used in observational real-world studies to evaluate patients mortality and their end-of-life morbidities and care [4]. For this reason, subject to Medicines and Healthcare products Regulatory Agency Independent Scientific Advisory Committee [5] approval, the Office for National Statistics (ONS) death registry [6] is one of the data sources that can be linked to GOLD at patient level for England. Access to both is provided by CPRD for those patients whose practices agreed to participate to the linkage process and that did not opt out. Hospital Episode Statistics (HES) data [7] can also be linked to GOLD for patients in England within the linkage schema and report death information for those who died in hospital. Because more than 50% of deaths in the U.K. do not occur in hospital and therefore cannot be captured in HES data [8], ONS linked death data are preferred over HES linked data as source of information on mortality.

When a study can access both GOLD and linked ONS death data, it is crucial for researchers to make consistent and meaningful use of mortality information coming from both datasets in a way that maximises their potential and minimises their weaknesses.

While causes of death are not considered to be well recorded in GOLD, dates of death instead are regarded to have very high accuracy [8]. On the other hand, ONS data are considered the gold standard for mortality records, but these linked data come with some limitations, including linkage coverage period, linkage correctness likelihood and linkage regional restrictions. Mortality dates can be present in only one, or both, of the two datasets. When date of deaths are provided in both, but are inconsistent, data quality is challenged [4]. This issue may come from incorrect data recording in one or both datasets or from a linkage mismatch. Indeed, the linkage between the two datasets comes with a match rank, which is a measure of match likelihood: if the linkage for a patient is not considered reliable or the record not acceptable, not only the ONS date of death should be disregarded, but also the accompanying causes of death. To our knowledge there are no guidelines on how to manage all the branches of this complex scenario in order to confidently and consistently use mortality data from GOLD and linked ONS. We have developed these missing recommendations and implemented them in a deterministic SQL algorithm that negotiates between GOLD and linked ONS mortality data to identify the most reliable dates of death. Our overarching aim is to improve the quality [9], reproducibility [10], transparency [11] and comparison [12] of real-world observational research when using mortality data based on GOLD and linked ONS.

2. Related work and Material

2.1. CPRD GOLD Death Data

CPRD is part of the U.K. Department of Health and Social Care, jointly sponsored by the Medicines and Healthcare products Regulatory Agency (MHRA) and the National Institute for Health Research (NIHR). It homes two anonymized longitudinal datasets whose data are collected via different software used by U.K. General Practitioners (GPs):

GOLD, established in 1987, derives from Vision [13] and the recently born Aurum, launched in October 2017 [14] derives from Emis Web [15]. Although GOLD and Aurum data are collected for practice and patient management purposes, the secondary use of clinical data in medical research is widely accepted when the main issues of accuracy, consistency and completeness of data are addressed [16].

The history of GOLD has been widely documented [17,18,19,20] and the dataset is recognised to be extremely influential with at present more than 2,300 peer-reviewed research publications based on its data [2]. Vision, the software collecting its data mentioned above, covers 9% of the market in England [21]. This system provides different ways of recording the death date of a patient, which is a problematic choice because by not following the Relational Model [22] for database management, it allows for data discrepancy and data redundancy. In order to manage these issues and to identify the best estimate for the date of death in GOLD, CPRD has developed an in-house algorithm [8], which reasons on these multiple sources within Vision and estimates the official GOLD *deathdate* field present in the *Patient* table. This estimated date of death has been recorded in GOLD September release for nearly 1.4 million people, corresponding to around 8% of the total number of historical and current acceptable patients. The CPRD Death Algorithm has proved to work well and, when the exact date of death is not important, it produces an estimate considered to be adequate for the majority of studies. In a recent validation against the ONS linked death registry [8] based on a random sample of GOLD patients from England, eligible for ONS linkage and with a recorded ONS death, more than 98% had death also reported in GOLD. There was exact date agreement in nearly 70% across the whole study period with an overtime increase from 53.4% in 1998 to 78.0% in 2013 when 98.8% of dates of death were in agreement within ±30 days.

2.2. ONS Death Data

ONS death data are considered the gold standard because they are based on official Medical Certificates of Cause of Death [23], which is a legal requirement and responsibility of the General Register Office for England and Wales in partnership with the Local Registration Service. Although the ONS death data are considered more accurate than GOLD death records, there are several strengths and weaknesses that need to be pondered:

1. ONS mortality data are available for linkage only for deaths registered in England. This means that GOLD is the only source of death information for patients from Wales, Scotland and Northern Ireland and for those in England registered with practices out of the linkage schema or that opted out.
2. ONS mortality data available for linkage are limited to a coverage period. This means than GOLD is the only source of mortality information for patients who died outside that time window.
3. In 2011 it has been calculated that deaths considered unexpected, accidental or suspicious are not registered in ONS within the expected 5-day period. This delay may account for a few days, but may also extend up to years for 5% of deaths [24]. This means that some recent deaths can be missing in ONS data, while potentially present in GOLD.
4. For a small number of patients ONS death dates are incomplete. This means that these death dates are not fully available in ONS, but they might be in GOLD.

2.3. GOLD-ONS Linkage

With the growing availability and use of routinely collected big clinical data, linkage of digital records from different sources has become increasingly engaged and influential

in medical research [25,26]. GOLD patients registered in England are eligible for ONS linkage if their GP practices enrolled into the schema and they did not opt out. For each linkage set released CPRD provides:

1. A linkage file for all GOLD patients in England reporting patient-by-patient eligibility flags for all the available linkages. Researchers can use this file to identify which patients are eligible for ONS death linkage, which in the GOLD September 2019 release based on linkage set 17 were 46.12% of all acceptable patients.
2. A coverage file containing information about the time period covered by the linkage. When ONS mortality data report some deaths before or after the coverage period, these data should be discarded as mistaken, including the causes of death. For example the current linkage set has coverage starting on 02/01/1998 and ending on 14/01/2019. Because the September 2019 GOLD release contains data up-to August 2019, GOLD is the only source of death information for patients who died before 1998 or between the 14th of February 2018 and August 2019 approximately.

The actual linkage between GOLD and ONS is then undertaken on behalf of CPRD by NHS Digital [27] as a trusted third party so that no patient personal information is held by CPRD or by researchers. The linkage is based on an 8-step deterministic algorithm, which looks out for matches of the NHS number, date of birth, gender and postcode, as shown in Table 1.

Step	NHS Number	Date of Birth	Gender	Postcode
1	Exact	Exact	Exact	Exact
2	Exact	Exact	Exact	
3	Exact	Partial	Exact	Exact
4	Exact	Partial	Exact	
5	Exact			Exact
6		Exact	Exact	Exact
7		Exact	Exact	Exact
8	Exact			

Table 1

The match rank reported in the ONS linked data indicates the step of the algorithm at which the match between a patient in GOLD and ONS was established. CPRD releases data only for match ranks lower or equal to 5, although higher match rank links are available on demand. The lower the match rank value, the stronger the likelihood for a correct match. The very existence of the match rank suggests that these linkages are not guaranteed and come with different level of confidence: for a minority of patients the linkage might be mistaken and the ONS linked data may not belong to the correct patients. Also, a minority of patients are linked to multiple death records and, although they have been removed in recent years before release, their data are available on request.

The most sophisticated piece of information of the four used by this 8-step algorithm is the NHS number, a 10-digit long identifier uniquely and consistently associated to patients across the health and social care system in England, Wales and Isle of Man. To minimise the risk of its incorrect recording (e.g. accidentally transposing or mistyping) the NHS number includes an error-detecting checksum digit in 10th position, which is generated using a variation of the International Standard Book Number (ISBN) modulus 11 algorithm [28]. The total accuracy of this algorithm is around 91%, leaving nearly 9% of double errors undetected [29]. Moreover, an NHS number may be recorded against the wrong patient and this is another reason why also other information is necessary to link patient records, such as the date of birth, the gender and the postcode.

3. Methods and Results

We have developed our guidelines on the assumption that GOLD death dates in practices that joined the ONS linkage program in England are adequate for the majority of studies, as recently proved [8]. We have also assumed that the quality of data in U.K. practices outside

England or inside but not eligible for ONS linkage, is on avarage comparable to that of English linkable practices. This assumption is reasonable considering that all GOLD data come only from practices that collect clinical data using the same software (i.e. Vision) and whose data quality is routinely evaluated as suitable for clinical research starting from an up-to-standard date [30]. Our methodology liaises between the dates of death provided by GOLD and ONS using linkage coverage period, linkage match rank and linkage regional limitations.

3.1. The guidelines

We have developed a set of comprehensive guidelines covering all possible scenarios when ONS mortality data are linked to GOLD. The first four recommendations are the most commonly used, while the last two cover rare cases where multiple ONS death records are associated to the same patient identifier. We have included also the latter because some studies could benefit from them (e.g. mother/baby studies) and the records are available from CPRD on request.

1. *If only GOLD death date is present, always use it as a reliable piece of mortality information because its validity has been verified*.

   This guideline makes sure that patients ineligible for linkage with ONS, which in the current linkage set are 25.11% from England and 28.77% from Scotland, Wales and Northern Ireland, have dates of death extracted from GOLD.

2. *If only ONS mortality information is present, use its date and causes of death if and only if the linkage match rank is equal to or lower than 2 and the date of death is within the linkage coverage period*.

   We have chosen to use only match rank 1 and 2 of the 5 provided by CPRD because we consider the exact date of birth, which is not required for match ranks from 3 to 5, essential information to identify an individual. We have been flexible about the

postcode, which is not required in match rank 2, because this information is the only one likely to change over time. This guideline ensures that of the 98.2% linkages provided by CPRD in the latest linkage set, only the most certain links are used, which account for 95.4%, while the others 2.8% are disregarded. This recommendation also makes sure that ONS mortality data referring to an untrustworthy time period are rejected.

3. *If both GOLD and ONS dates of death are available, use ONS date and causes of death if the ONS death date is within the linkage coverage period and the match rank is equal to 1 or 2*.

   This guideline guarantees that we use the most reliable mortality data available for the 46.12% of GOLD patients eligible for ONS linkage.

4. *If both GOLD and ONS are present, and ONS date of death is outside the linkage coverage period or the match rank is greater or equal to 3, use GOLD date of death to validate ONS: if the two dates coincide use ONS date and causes of death, otherwise use GOLD date of death only*.

   This guideline assures that no reliable mortality information is dismissed even if outside the linkage coverage period or outside the 1-2 match rank range.

5. *If there are more ONS records for the same patient with death dates within the linkage coverage period, match ranks equal to 1 or 2 and GOLD date of death is not available, always choose the ONS record with the lower match rank, if any, otherwise disregard ONS mortality data*.

   This guideline makes sure the linkage with the highest likelihood is always chosen and that ONS mortality data are rejected if no reasonable choice can be made.

6. *If there are more ONS records for the same patient and GOLD date of death is available, use it to validate ONS mortality data*.

- *If GOLD death date coincides with one of the ONS records, use the latter without checking the linkage coverage and match rank. If GOLD coincides with more than one ONS death dates, use GOLD date and disregard ONS records.*
- *If GOLD date of death does not coincide with any of the ONS ones, test if one of the ONS dates of death falls within 30 days from the GOLD one and use it if and only if it is within the linkage coverage period and match rank is 1 or 2. Otherwise disregard ONS mortality data and use GOLD date of death.*

This guideline assures we combine all the mortality data available to make informed decisions and that ONS death information is dismissed only when there is lack of evidence it can be trusted.

3.2. The SQL algorithm

We have implemented these guidelines in an SQL algorithm based on GOLD data specification version 2.0 [30] and on linkage set 17 data documentation version 2.0 for ONS [31], which are the most recent at the time of writing. Developed in MySQL 5.7.24 [32] using InnoDB engine, it assumes that GOLD Patient and ONS death data have been imported without further processing into homonymous tables. The algorithm allows for customisation of the match rank threshold (i.e. @max_death_matchrank) and of the time windows allowed between GOLD and ONS death dates in case of single (i.e. @max_days_single) and multiple ONS linkages (i.e. @max_days_multi), as our choices are conservative and could not be suitable for some studies.

# Customise variables SET @max_death_matchrank = 2; SET @max_days_single = 0; SET @max_days_multi = 30;	branch
# Read ONS coverage dates SET @ons_start = (SELECT start FROM linkage_coverage WHERE data_source = 'ons_death'); SET @ons_end = (SELECT end FROM linkage_coverage WHERE data_source = 'ons_death');	branch
# Reconcile GOLD.deathdate WITH ONS.dod	

```
SELECT goldp.patid,
CASE
    WHEN t.ons_num IS NULL THEN      # No ONS death records are available
        CASE
            WHEN goldp.deathdate IS NOT NULL THEN goldp.deathdate                    1
            ELSE NULL                                                                2
        END
    WHEN t.ons_num = 1 THEN          # 1 ONS death record is available
        CASE
            WHEN goldp.deathdate IS NULL THEN
                CASE
                    WHEN (ons.death_matchrank <= @max_death_matchrank
                    AND ons.dod >= @ons_start AND ons.dod <= @ons_end)
                    THEN ons.dod                                                     3
                    ELSE NULL                                                        4
                END
            WHEN ABS(DATEDIFF(ons.dod,goldp.deathdate)) <= @max_days_single
            THEN ons.dod                                                             5
            ELSE
                CASE
                    WHEN (ons.death_matchrank <= @max_death_matchrank
                    AND ons.dod >= @ons_start AND ons.dod <= @ons_end)
                    THEN ons.dod                                                     6
                    ELSE goldp.deathdate                                             7
                END
        END
    WHEN t.ons_num > 1 THEN          # > 1 ONS death records are available
        CASE
            WHEN goldp.deathdate IS NULL THEN  # No GOLD deathdate is available
                CASE
                    WHEN (SELECT num_record FROM (SELECT patid, death_matchrank,
                    COUNT(*) as num_record
                    FROM ons
                    WHERE (death_matchrank <= @max_death_matchrank
                    AND dod >= @ons_start AND dod <= @ons_end)
                    GROUP BY patid, death_matchrank) as t2
                    WHERE patid = goldp.patid
                    ORDER BY death_matchrank ASC LIMIT 1) = 1 THEN
                    (SELECT dod  FROM ons
                    WHERE patid = goldp.patid
                    AND (death_matchrank <= @max_death_matchrank
                    AND dod >= @ons_start AND dod <= @ons_end)
                    ORDER BY death_matchrank ASC LIMIT 1)                            8
                    ELSE NULL                                                        9
                END
            WHEN goldp.deathdate in (SELECT dod FROM ons WHERE patid = goldp.patid)
            THEN   # GOLD deathdate coincides with one of the ONS dods
                CASE                                                                 branch
```

WHEN (SELECT num_record FROM (SELECT patid, dod, death_matchrank, COUNT(*) as num_record FROM ons GROUP BY patid, dod, death_matchrank) as t2 WHERE patid = goldp.patid AND dod = goldp.deathdate ORDER BY death_matchrank ASC LIMIT 1) = 1 THEN (SELECT dod FROM ons WHERE patid = goldp.patid AND dod = goldp.deathdate ORDER BY death_matchrank ASC LIMIT 1)	10
ELSE goldp.deathdate	11
END	
ELSE # one of the acceptable ONS dods is within 30 days of GOLD deathdate	
CASE	
WHEN (SELECT num_record FROM (SELECT patid, dod, death_matchrank, COUNT(*) as num_record FROM ons WHERE (death_matchrank <= @max_death_matchrank AND dod >= @ons_start AND dod <= @ons_end) GROUP BY patid, dod, death_matchrank) as t2 WHERE patid = goldp.patid AND ABS(DATEDIFF(dod,goldp.deathdate)) <= @max_days_multi ORDER BY death_matchrank ASC LIMIT 1) = 1 THEN (SELECT dod FROM ons WHERE patid = goldp.patid AND ABS(DATEDIFF(dod,goldp.deathdate)) <= @max_days_multi AND (death_matchrank <= @max_death_matchrank AND dod >= @ons_start AND dod <= @ons_end) ORDER BY death_matchrank ASC LIMIT 1)	12
ELSE goldp.deathdate	13
END	
END	
END as dod_final	
FROM patient as goldp	
LEFT JOIN ons on goldp.patid = ons.patid	
LEFT JOIN (SELECT patid, COUNT(*) as ons_num FROM ons GROUP BY patid) as t on t.patid = goldp.patid;	

Table 2

Table 3 reports examples of decision making for the first 7 branches of the algorithm on test data created to synthesize all the most frequent scenarios represented by guidelines 1 to 4. The algorithm can be easily expanded to accommodate ONS linked primary and/or secondary causes of death using the same logic and the *ons.causes* column shows when ONS causes of death are safe to be used.

patid	patient deathdate	ons match_rank	ons dod	dod_final	ons causes	explanation	branch
patid1	2001-10-18			2001-10-18		Only GOLD date present	1
patid2						None present	2
patid3		2	2006-06-29	2006-06-29	yes	Valid ONS record present	3
patid4		1	1995-10-03			ONS date out of coverage	4
patid5	2003-07-05	1	2003-07-05	2003-07-05	yes	Same date, use ONS	5
patid6	2008-01-02	1	2009-01-02	2009-01-02	yes	Valid ONS record present	6
patid7	2005-01-20	4	2013-02-15	2005-01-20		Unacceptable ONS match_rank	7

Table 3

Table 4 reports examples of decision making performed by this second part of the algorithm on test data created to synthesize all the possibilities when multiple ONS records associated to the same patient identifier might be present. These rare scenarios are covered by guidelines 5 and 6.

patid	patient deathdate	ons match_rank	ons dod	dod_final	ons causes	explanation	branch
patid8		1 2	2011-10-09 2009-01-09	2011-10-09	yes	Chosen ONS record with the lower match_rank	8
patid9		1 1	2002-05-04 2004-08-18			Impossible to choose between ONS records: same match rank	9
patid10	2009-12-22	1 1 2	1998-10-23 1999-03-31 2009-12-22	2009-12-22	yes	Chosen ONS record with date equal to GOLD date	10
patid11	2004-03-02	2 2	2004-03-02 2004-03-02	2004-03-02		Chosen GOLD date, impossible to choose between ONS records	11
patid12	2009-01-30	1 1	2009-01-29 2010-07-30	2009-01-29	yes	Chosen ONS record with date within 30 days of GOLD date	12
patid13	2003-02-17	3 4	2005-03-15 2012-06-11	2003-02-17		Chosen GOLD date, unacceptable ONS match_ranks	13

Table 4

The algorithm is deterministic and has been tested by checking the calculated date of death for each of the 13 branches of the decision tree. The execution plan of the algorithm has been analysed using MySQL EXPLAIN EXTENDED statement. Assuming

that the two tables have an index on the joined columns, the estimated time complexity is linear and proportional to the tables' dimensions. In particular, if N and M are the numbers of rows in *patient* and *ons* tables respectively and P is the number of deceased patients present in *ons* table with P<=M and P<=N, the maximum time complexity of the algorithm is $O(N+4*M)$, where the factor 4 reflects the number of derived queries. This value could be lowered by creating temporary tables, however we decided that this performance was more than acceptable considering that for the majority of studies M and P will be both lower than 25% of N.

This algorithm could be quickly adapted to any other SQL dialect and/or be embedded in programs or scripts written in any high level programming language, (e.g. Python, C++, PHP, etc.) and/or executed in statistical packages (e.g. SAS, R, Stata, SPSS).

4. Conclusion

We have identified a lack of guidance on how to manage the challenges presented when ONS mortality data are linked to GOLD dataset, including linkage time coverage, linkage correctness likelihood, linkage regional limitations and data discrepancy. We have therefore produced novel reconciling guidelines for researchers to address all these issues and developed an SQL algorithm, which implements them. This algorithm should be easily imported and utilised in a variety of computing and statistical settings. This automation should help to improve quality, reproducibility and transparency of clinical research and comparison between studies.

Funding

The research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Contributors

AD and DPA contributed to the death dates reconciling guidelines, AD led its development and implementation. All authors contributed to the writing and editing of the paper.

Competing interests

The authors have no competing interests to declare.

Summary points

What is already known?

- GOLD and ONS mortality datasets can be linked for those patients registered with practices in England that agreed to take part to the linkage process and did not opt out
- ONS death data are considered the gold standard on mortality
- GOLD death dates have been recently assessed against ONS linked death dates and their quality deemed sufficient for the majority of epidemiological studies

What this study has added?

- Reconciling guidelines on how to manage the linkage coverage period, linkage correctness likelihood, linkage regional limitation and data discrepancies between GOLD and ONS mortality datasets
- An SQL algorithm that implements our recommendations with the aim to improve clinical research quality, reproducibility, transparency and comparison.

References

[1] (2019, September) CPRD. [Online]. https://www.cprd.com/primary-care

[2] (2019, September) CPRD - Bibliography. [Online]. https://www.cprd.com/bibliography

[3] (2019, September) CPRD - GOLD Release Notes. [Online]. https://cprdcw.cprd.com/_docs/Release_Notes_September2019.pdf

[4] Amelia Harshfield, Gary A Abel, Stephen Barclay, and Rupert A Payne, "Do GPs accurately record date of death?," *BMJ Supportive & Palliative Care*, 2018.

[5] (2019, September) Independent Scientific Advisory Committee for MHRA database research. [Online]. https://www.gov.uk/government/groups/independent-scientific-advisory-committee-for-mhra-database-research

[6] (2019, September) Office for National Statistics - Death Registration data. [Online]. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths

[7] (2019, September) NHS Digital - Hospital Episode Statistics. [Online]. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics

[8] Arlene M Gallagher, Daniel Dedman, Shivani Padmanabhan, Hubert G M Leufkens, and Frank de Vries, "The accuracy of date of death recording in the Clinical Practice Research Datalink GOLD database in England compared with the Office for National Statistics death registrations," *Pharmacoepidemiol Drug Saf*, pp. 1-7, 2019.

[9] Wenfei Fan, "Data Quality: Theory and Practice," in *Web-Age Information management (WAIM)*, Harbin, China, 2012.

[10] Roger D Peng, "Reproducible Research in Computational Science," *Science*, vol. 334, no. 6060, pp. 1226-1227, December 2011.

[11] Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biology*, vol. 11, no. 8:R86, pp. 1-13, 2010.

[12] Andrew Gelman and Eric Loken, "The Statistical Crisis in Science," *American Scientist*, vol. 102, no. 6, pp. 460-465, November-December 2014.

[13] Vision. (2019, September) Vision Health. [Online]. https://www.visionhealth.co.uk/

[14] Achim Wolf et al., "Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum," *International Journal of Epidemiology*, pp. 1-8, 2019.

[15] (2019, September) EMIS Health. [Online]. https://www.emishealth.com/

[16] Ioana Danciu et al., "Secondary use of clinical data: The Vanderbilt approach," *Journal of Biomedical Informatics*, vol. 52, pp. 28-35, 2014.

[17] D H Lawson, V Sherman, and Jennifer Hollowell, "The General Practice Research Database," *Quarterly Journal of Medicine (QJM)*, vol. 91, pp. 445–452, 1998.

[18] Louise Wood and Carlos Martinez, "The General Practice Research Database," *Drug Safety*, vol. 27, no. 12, pp. 871-881, 2004.

[19] John Parkinson, Sarah Davis, and Tjeerd van Staa, "The General Practice Research Database: Now and the Future," in *Pharmacovigilance*, Ronald D Mann and Elizabeth B Andrews, Eds.: John Wiley & Sons, Ltd, 2007, ch. 27, pp. 341-348.

[20] Louise Wood, "GPRD in the UK," in *Pharmacovigilance*, R D Mann and E B Andrews, Eds.: John Wiley & Sons, 2002, ch. 29, pp. 373-378.

[21] Evangelos Kontopantelis et al., "Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study," *BMJ Open*, vol. 8, p. e020738, 2018.

[22] Edgar Frank Codd, "A Relational Model of Data for Large Shared Data Banks," *Communications of the ACM*, vol. 13, no. 6, pp. 377-387, 1970.

[23] (2019, September) Completing a medical certificate of cause of death (MCCD). [Online]. https://www.gov.uk/government/publications/guidance-notes-for-completing-a-medical-certificate-of-cause-of-death

[24] (2019, September) Office for National Statistics - Impact of registration delays on mortality statistics. [Online]. https://webarchive.nationalarchives.gov.uk/20160106020016/http://www.ons.gov.uk/ons/guide-method/user-guidance/health-and-life-events/impact-of-registration-delays-on-mortality-statistics/index.html

[25] Ruth Gilbert et al., "GUILD: GUidance for Information about Linking Data sets," *Journal of Public Health (Oxford)*, vol. 40, no. 1, pp. 191-198, Mar 2018.

[26] Shivani Padmanabhan et al., "Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications," *European Journal of Epidemiology*, vol. 34, pp. 91-99, 2019.

[27] (2019, September) NHS Digital. [Online]. https://digital.nhs.uk/

[28] Andy Boyd, Richard Thomas, and John Macleod, "NHS Number and the Systems Used to Manage Them: An Overview for Research Users," *Cohort & Longitudinal Studies Enhancement Resources*, pp. 1-26, April 2018.

[29] Aravindan Siddharth, "Error Detection in Numeric Codes," *Resonance*, vol. 17, no. 7, pp. 653-671, July 2012.

[30] (2019, September) CPRD - GOLD Data Specification. [Online]. https://cprdcw.cprd.com/_docs/CPRD_GOLD_Full_Data_Specification_v2.0.pdf

[31] (2019, September) CPRD - ONS Data Documentation. [Online]. https://cprdcw.cprd.com/_docs/Documentation_Death_set17_v2.0.pdf

[32] (2019, September) MySQL. [Online]. https://dev.mysql.com/