Journal of Big Data

RESEARCH Open Access

Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test

Pradeep S. Virdee[1*], Alice Fuller[2], Michael Jacobs[3], Tim Holt[2] and Jacqueline Birks[1]

*Correspondence:
pradeep.virdee@csm.ox.ac.uk
[1] Centre for Statistics in Medicine, Botnar Research Centre, Nuffield Orthopaedic Centre, NDORMS, University of Oxford, Windmill Road, Oxford OX3 7LD, UK
Full list of author information is available at the end of the article

Abstract

A Full Blood Count (FBC) is a common blood test including 20 parameters, such as haemoglobin and platelets. FBCs from Electronic Health Record (EHR) databases provide a large sample of anonymised individual patient data and are increasingly used in research. We describe the quality of the FBC data in one EHR. The Test dataset from the Clinical Research Practice Datalink (CPRD) was accessed, which contains results of tests performed in primary care, such as FBC blood tests. Medical codes and entity codes, two coding systems used within CPRD to identify FBC records, were compared, with levels of mismatched coding, and number that could be rectified reported. The reliability of units of measurement are also described and missing data discussed. There were 14 entity codes and 138 medical codes for the FBC in the data. Medical and entity codes consistently corresponded to the same FBC parameter in 95.2% (n = 217,752,448) of parameters. In the 4.8% (n = 10,955,006) mismatches, the most common parameter rectified was mean platelet volume (n = 2,041,360) and 1,191,540 could not be rectified and were removed. Units of measurement were often either missing, partially entered, or did not appear to correspond to the blood value. The final dataset contained 16,537,017 FBC tests. Applying mathematical equations to derive some missing parameters in these FBCs resulted in 15 of 20 parameters available per FBC on average, with 0.3% of FBCs having all 20 parameters. Performing data quality checks can help to understand the extent of any issues in the dataset. We emphasise balancing large sample sizes with reliability of the data.

Keywords: Clinical practice research datalink, Full blood count, Blood test, Data quality, Data validation

Introduction

Electronic Health Records (EHR) are databases that store routinely-collected anonymised individual patient data. These databases were set up to aid patient care and monitor clinical services. Their use for research is a secondary development, which has allowed researchers to conduct large-scale medical studies and perform retrospective

longitudinal analyses with large sample sizes. The use of EHR for research is becoming increasingly common over time.

Clinical Practice Research Datalink

Where do the data come from?

EHRs are designed to improve efficiency in practices, such as to manage appointments and provide patient services, such as ordering repeat prescriptions. Under the GP Systems of Choice (GPSoC) framework [1], primary care practices use an EHR software system that best suits their data management needs. The Clinical Practice Research Datalink (CPRD) [2] collects patient records from contributing GP practices in the UK using the Vision and, more recently, EMIS software systems.

Anonymised data are extracted by the CPRD through regular downloads, a process that does not require active intervention by the practices. Over 600 UK practices contribute to CPRD. The data encompass 42 million patient lives from over the last 30 years. A full description of CPRD can be found on the CPRD website [3].

CPRD structure

The CPRD database is formed of 10 main datasets, as described in the CPRD Data Specification [4]. Patient records are coded using medical codes, which are the numeric equivalent of Read codes available in patient records from the GP system [5, 6]. Other coding systems, such as the International Classification of Diseases-10 (ICD-10) codes to identify diseases and malignancies, are available in linked EHRs only [7].

One of the 10 datasets, called the Test dataset, holds records of tests and examinations performed in primary care, including laboratory tests. CPRD refers to each type of each test as an entity, assigning each a unique number, the entity code. For example, haemoglobin and platelet count tested as part of a Full Blood Count (FBC) blood test are assigned a unique entity code of 173 and 189, respectively. The entity code is generated by the EHR system but is not visible to practice staff—it is used internally by the CPRD to provide a convenient way to file data that preserves (to some extent) the original data structure in practice. In the Test dataset, individual items in a patient's record are coded using both medical codes and entity codes.

The list of data items in the Test dataset include the pseudo-anonymised patient identification number, the medical and entity code corresponding to the test or examination performed, the date it was performed, and the test results.

Full blood count

A FBC is a blood test commonly ordered in both primary and secondary care in the UK. A single FBC test includes up to 20 individual parameters [8], although additional parameters may sometimes be measured. A patient's blood sample, labelled using their name and NHS number, is delivered to a haematology laboratory and run through a processing machine, referred to as an analyser. Analysers have been used for many decades to derive FBC values and not all parameters were historically always derived. In the last decade or so, analysers derive the values for all 20 FBC parameters. Nine parameters are measured directly from the blood sample: red blood cell count, white blood cell count, haemoglobin, platelet count, basophil count, eosinophil count, lymphocyte

count, monocyte count, and neutrophil count. The remaining 11 are parameters that describe the nine measured parameters, derived using mathematical formulae programmed into the analyser: mean platelet volume, haematocrit (or packed cell volume), mean corpuscular volume, mean corpuscular haemoglobin, mean corpuscular haemoglobin concentration, red blood cell distribution width, basophil percentage, eosinophil percentage, lymphocyte percentage, monocyte percentage, and neutrophil percentage. Units of measurement used in practice have changed over time. An FBC report includes the resulting blood levels, units of measurement, and assigned medical codes for each FBC parameter.

The FBC report, labelled with the patient's name and NHS number, is electronically returned to the practice, where it is electronically assigned to the patient's electronic record, a process that is largely automated. It is then examined and filed by a clinician who will decide on necessary actions. This report contributes to the CPRD database at its next download, with each parameter assigned a medical code from the laboratory and entity code from the EHR system.

Data quality check aims

Laboratory data, including the FBC, from EHR databases are commonly used in research studies [9–14]. Although data cleaning is a common form of data preparation before analysis in practice, data quality assessments and validation are often not performed. A systematic review identified many barriers to perfoming quality checks, including large amounts of unstructured data, challenges with patient identification and matching, problems with data extraction, and unfamiliarity with data quality assessment [15]. However, data quality checks are a crucial step to assess representativeness of clinical practices and ensure reliability of the results of any analyses. In one systematic review, all individual studies agreed that data accuracy and data completeness were key factors to consider when designing EHR studies [16]. A second review highlighted a need for a generalised approach to assess EHR data quality [17].

The aim of this study was to report a methodological approach to assess the quality of laboratory data from CPRD, demonstrated with application to the FBC. Our recent systematic review has identified many studies that use FBC data (n = 512), with 4% of 53 eligible studies using FBC data performing data validation before analysis [18–21]. As laboratory data are frequently used across medical research, we provide recommendations and guidance for researchers who wish to access and analyse EHR data in the future, and make available our statistical coding used to perform the data validation of CPRD, which other researchers can make use of.

Methods

CPRD data was accessed for a study period of 1st January 2000 to 28th April 2015 (data cut date) and approved by the CPRD Independent Scientific Advisory Committee, which covers ethical approval (14_195RMn2A2R). Patients aged at least 40 years at study entry with at least one FBC blood test in the Test dataset were included in the analysis because FBCs are more commonly performed in this age group.

A flow chart of our approach to assess the quality of the laboratory test data in CPRD is provided in Fig. 1.

Fig. 1 Flow chart of the methodological approach to assess the quality of laboratory test data in CPRD

FBC-related codes

The Test dataset was actively searched to identify FBC-related medical codes and entity codes. The derived code list was compared to independent lists from relevant published studies and clinical code repositories [22–24] to validate the list.

Medical and entity code comparison

The medical and entity code assigned to each parameter was compared by checking the types of medical codes assigned to each FBC-related entity code and vice-versa. Parameters were considered to be consistently coded if the medical and entity code corresponded to the same FBC parameter, and mismatched if otherwise.

Six FBC parameters do not have their own entity code (mean platelet volume, basophil proportion, eosinophil proportion, lymphocyte proportion, monocyte proportion, and neutrophil proportion), as indicated in the CPRD entity code dictionary. Therefore, we stratified mismatches into 3 strata:

1. Where one code suggested a particular FBC parameter but the other suggested a different FBC parameter, with existing medical or entity codes for both parameters that could have been assigned.
2. Where the medical code suggested one of the six FBC parameters without an existing entity code. They were considered inconsistent because they were assigned an entity code for a different FBC parameter by the EHR.
3. Where one code suggested a particular FBC parameter but the other suggested the record was not FBC-related.

Availablility of mismatched parameters

In mismatched pairs, the parameter suggested by either code was checked to see if it was already available for that FBC test (an FBC can include 20 individual parameters). For example, if one code suggested haemoglobin and the other suggested neutrophil count, it was checked whether haemoglobin and neutrophil count were already available in the same FBC. The three strata were analysed separately.

Rectify mismatches

The blood values and corresponding units of mismatched pairs were used to classify each as one of the two parameters suggested by either code, depending on which they best reflected. Consideration was given to possible plausibility of values for each

suggested blood parameter that could be used to differentiate between the two suggested parameters. Parameters that could not be rectified were removed.

Standardising FBC units

In the resulting dataset of consistently coded or rectified parameters, standardisation of the blood values to a single, conventional unit of measurement was planned. Parameters with unreliable units, such as those partially entered or where the value and unit did not appear to match, were deleted.

To identify the extent of any extreme or implausible blood values, each parameter was converted into quantiles and the mean, median, and range for each calculated. All parameters were divided into 10 quantiles (or deciles), except basophils, eosinophils, lymphocytes, monocytes, and neutrophils, which were divided into four quantiles (or quartiles) because the range of possible values is very small such that deciles could not be derived.

Derive missing FBC values

Some FBC parameters are mathematically related so missing FBC values can be derived using known values. The blood test date, which was the only available indicator to separate each FBC if a patient had multiple, was initially examined to ensure we derived values in a single FBC test using other values within that same FBC.

Subsequently, 25 known mathematical equations were applied to derive missing values, where possible (see Additional file 1: 1). Equations for haematocrit, red blood cell distribution width, and mean platelet volume exist, but rely on information not available in the CPRD dataset and could not be used. Deriving a parameter's value meant that it was available for use in an equation for another parameter, so we recursively applied the 25 equations until no further values could be derived.

Describe FBC data

The final dataset was summarised after all amendments, including the number of parameters and FBCs available, extreme or implausible blood values, and missing data.

Statistical analysis

A descriptive analysis was performed, with continuous variables described using mean with Standard Deviation (SD) or median with range and categorical variables described using counts and proportions. We used Stata 15.1 for all analyses.

Results

The CPRD Test dataset contained 695,139,617 test or examination records from 658 primary care practices.

FBC-related codes

In total, there were 325 different entity codes and 10,963 different medical codes used in the Test dataset. Table 1 shows a list of medical and entity codes related to the FBC we identified from the Test dataset. These codes were consistent with existing lists [22–24]. Codes in our list resulted in 228,707,454 FBC parameters among 2,914,589 patients.

Table 1 Entity and medical codes in CPRD corresponding to the FBC

FBC parameter	Entity code	Entity term	Medical code	Medical term
Red blood cell count	194	Red blood cell count	17	Red blood cell (RBC) count
			13,788	Nucleated red blood cell count
			26,931	RBC count NOS
			26,932	RBC count normal
			26,933	RBC count low
			44,213	RBC count abnormal
			50,182	Red cell mass
			57,136	RBC count raised
			58,853	RBC count borderline low
			70,079	RBC count borderline raised
White blood cell count	207	Total White Blood cell count	15	Total white cell count
			1955	Leucopenia
			3372	Leucopenia
			4760	Leucocytosis
			4996	White cell count abnormal
			13,817	White blood count
			13,818	White cell count
			18,516	Leucocytosis
			22,293	Leucocytosis -high white count
			26,325	Leucopenia—low white count
			26,946	Total white cell count NOS
			26,947	Total white blood count
			26,948	White cell count normal
			38,198	Leucocyte count
			45,115	Diff. white cell count normal
			48,015	Total WBC (IMM)
			48,341	Polymorphonuclear leukocyte count
			53,865	Leucocytosis
			92,372	Leukocytosis

Table 1 (continued)

FBC parameter	Entity code	Entity term	Medical code	Medical term
Haemoglobin	173	Haemoglobin	4	Haemoglobin estimation
			739	Anaemia unspecified
			795	Iron deficiency anaemias
			3942	Haemoglobin low
			10,404	Hb estimation
			13,755	Haemoglobin—sample sent
			26,272	Haemoglobin borderline high
			26,908	Haemoglobin abnormal
			26,909	Haemoglobin normal
			26,910	Haemoglobin borderline low
			26,912	Haemoglobin estimation NOS
			26,913	Haemoglobin high
			33,284	Haemoglobin requested
			35,749	Haemoglobin very low
			39,601	Haemoglobin not estimated
			41,531	Haemoglobin very high
Haematocrit/packed cell volume	312	Packed cell volume	40	Haematocrit
			99	Haematocrit—PCV
			14,240	Packed cell volume
			19,836	Packed cell volume—PCV
			23,476	Haematocrit—PCV—high
			27,143	Haematocrit—PCV—NOS
			27,144	Haematocrit—PCV—normal
			27,145	Haematocrit—PCV—low
			41,478	Haematocrit—PCV—abnormal
			55,365	Haematocrit—borderline low
			62,347	Haematocrit—borderline high
Mean corpuscular volume	182	Mean corpuscular volume	10	Mean corpuscular volume (MCV)
			2480	MCV—raised
			13,774	Mean cell volume
			26,920	MCV—NOS
			26,921	MCV—normal
			26,922	MCV—low
			41,160	MCV—borderline raised
			52,874	MCV—borderline low

Table 1 (continued)

FBC parameter	Entity code	Entity term	Medical code	Medical term
Mean corpuscular haemoglobin	180	Mean corpuscular haemoglobin	20	Mean corpusc. haemoglobin(MCH)
			23,214	Mean cell haemoglobin
			26,917	MCH—normal
			26,918	MCH—low
			40,170	MCH—NOS
			47,174	MCH—borderline raised
			49,225	MCH—abnormal
			51,616	MCH—borderline low
			61,951	MCH—raised
Mean corpuscular haemoglobin concentration	181	MCH Hb Concentration	30	Mean corpusc. Hb. conc. (MCHC)
			26,919	MCHC—raised
			39,202	MCHC—NOS
			47,345	MCHC—normal
			55,183	MCHC—low
			64,474	MCHC—borderline low
			72,488	MCHC—borderline raised
Red blood cell distribution width	361	RBC red blood cell size	64	Red blood cell distribution width
			1191	RBC's—anisocytosis
			1962	RBC's—macrocytic
			9933	RBC's—microcytic
			19,837	RBC—red blood cell size
			51,484	Red blood cell size normal
			52,050	Red blood cell size NOS
Platelets	189	Platelets	7	Platelet count
			3320	Thrombocythaemia
			4006	Thrombocytopenia
			4415	Platelet count abnormal
			26,926	Platelet count NOS
			26,927	Platelet count normal
			37,666	Platelet aggregation test
Mean platelet volume	–	–	14,166	Mean platelet volume
Basophil count	313	Basophil count	25	Basophil count
			27,146	Basophil count NOS
			27,147	Basophil count normal
			27,148	Basophil count abnormal
			53,404	Basophilia
Basophil proportion	–	–	14,096	Percentage basophils
Eosinophil count	168	Eosinophil count	22	Eosinophil count
			13,742	Eosinopenia
			18,531	Eosinophil count raised
			26,905	Eosinophil count NOS
			26,906	Eosinophil count normal
Eosinophil proportion	–	–	19,760	Percentage eosinophils

Table 1 (continued)

FBC parameter	Entity code	Entity term	Medical code	Medical term
Lymphocyte count	208	Lymphocyte count	19	Lymphocyte count
			3189	Lymphocytosis
			11,240	Lymphopenia
			23,120	Lymphocyte count normal
			23,121	Lymphocytosis—absolute
			26,949	Lymphocyte count NOS
			26,950	Lymphocyte count abnormal
			32,932	Reactive lymphocyte count
			34,551	Total lymphocyte count (IMM)
			37,677	Lymphocytosis—relative
			42,346	Abnormal lymphocytes
			74,019	Lymphocyte function test
Lymphocyte proportion	–	–	17,621	Percentage lymphocytes
Monocyte count	183	Monocyte count	21	Monocyte count
			9248	Monocytosis
			13,776	Monocyte count NOS
			26,923	Abnormal monocytes
			26,924	Monocyte count normal
			26,925	Monocyte count raised
			44,189	Monocyte count abnormal
			72,849	Monocytopenia
Monocyte proportion	–	–	13,775	Percentage monocytes
Neutrophil count	184	Neutrophil count	18	Neutrophil count
			4463	Neutropenia
			13,777	Neutrophil count NOS
			15,725	Neutrophilia
			23,112	Neutrophil count normal
			23,113	Neutrophil function test
			31,382	Neutrophil count abnormal
			105,211	Band neutrophil count
Neutrophil proportion	–	–	17,622	Percentage neutrophils

FBC full blood count

Medical and entity code comparison

All 228,707,454 FBC parameters had a medical and entity code assigned. Coding was consistent for 95.2% (n = 217,752,448) and mismatches occurred in 4.8% (n = 10,955,006) (see Table 2). Mean age at study entry was 42.3 years and 42.8 years for patients with consistently coded parameters and mismatched pairs and 44.6% and 43.3% were male, respectively. Mismatched pairs were among 94.1% (n = 619) of the 658 practices. There were 37.9% (n = 4,156,438) from FBCs performed within five years of the data cut (2010 to 2015) and the majority of tests with mismatches were performed between 2005 and 2010 (42.0%, n = 4,595,900).

Table 2 Number of full blood count parameters at different stages of the quality check

FBC parameter	1. In the CPRD Test dataset			2. After rectifying mismatches	3. After standardisation	4. After deriving values
	Consistent[a]	Mismatched[a]				
		Using entity codes	Using medical codes			
RBC	17,496,396	–	2233	17,496,431	14,483,956	14,989,332
WBC	17,510,110	4	84,959	17,510,410	16,098,918	13,391,463
Haemoglobin	18,129,224	425	3318	18,136,578	13,241,660	15,956,530
Haematocrit	16,229,494	–	90	16,229,502	3,901,766	14,929,672
MCV	16,972,637	3	62	16,972,661	16,190,073	15,508,062
MCH	15,828,539	48	2443	15,828,610	14,776,728	14,978,026
MCHC	11,600,421	–	99	11,600,434	7,872,690	14,758,965
RDW	7,187,721	–	92,334	7,187,769	319,775	312,052
Platelets	17,321,447	2,267,404	798	17,321,470	16,110,377	15,409,813
MPV	–	–	2,299,660	2,049,996	2,041,360	1,964,340
Basophil count	15,064,082	1,647,073	625	15,064,119	13,939,287	11,532,656
Basophil %	–	–	1,670,876	1,579,167	1,498,690	11,577,466
Eosinophil count	15,963,917	1,579,993	6578	15,963,935	14,761,804	11,609,390
Eosinophil %	–	–	1,601,225	1,516,946	1,505,596	11,619,039
Lymphocyte	16,164,725	1,634,084	6825	16,164,762	14,950,124	12,202,708
Lymphocyte %	–	–	1,616,430	1,531,345	1,508,370	12,195,026
Monocyte count	16,047,916	1,601,135	7117	16,047,958	14,687,897	12,068,644
Monocyte %	–	–	1,625,740	1,537,874	1,510,233	12,076,052
Neutrophil count	16,235,819	1,586,187	9391	16,235,860	15,061,611	12,306,548
Neutrophil %	–	–	1,609,549	1,540,087	1,521,541	12,299,431
Not FBC		638,650	314,654			
Total (% of 228,707,454)	217,752,448 (95.2%)	10,955,006 (4.8%)	10,955,006 (4.8%)	227,515,914 (99.5%)	185,982,456 (81.3%)	241,685,233 (105.7%)

FBC full blood count, *RBC* red blood cell count, *WBC* white blood cell count, *MCV* mean corpuscular volume, *MCH* mean corpuscular haemoglobin, *MCHC* mean corpuscular haemoglobin concentration, *RDW* red blood cell distribution width, *MPV* mean platelet volume

[a] Consistent records are those where the medical code and entity code correspond to the same parameter. Those where this does not apply were considered to have mismatched coding

Of the mismatched pairs, 44,349 had medical and entity codes correspond to different FBC parameters, where both suggested parameters had existing medical or entity codes that could have been assigned (strata 1). See Table 3 for further details. The most common mismatch was entity 208 (lymphocyte count) and medical code 38,189 (white blood cell count), with 43,869 occurrences.

There were 10,272,007 mismatched pairs because six do not have an existing entity code and an alternative assigned (strata 2). See Table 3 for further details. The most common mismatch was entity 189 (platelets) and medical code 14,166 (mean platelet volume), with 2,267,404 occurrences.

The remaining 638,650 mismatched pairs had one code suggest a FBC parameter but the other suggested the test was not from a FBC blood test (strata 3). There were 12 FBC entity codes where the corresponding medical code was not a FBC parameter,

Table 3 Number of mismatched pairs per FBC parameter

FBC parameter (medical code)	FBC parameter (entity code)	Number of mismatched pairs
Strata 1[a]:		
White blood cell count (38,198)	Lymphocyte count (208)	43,869
Neutrophil count (18)	Haemoglobin (173)	425
Mean corpuscular haemoglobin concentration (30)	Mean corpuscular haemoglobin (180)	48
Mean corpuscular volume (10)	White blood cell count (207)	4
Haemoglobin (795)	Mean corpuscular volume (182)	3
Strata 2[b]:		
Mean platelet volume (14,166)	Platelets (189)	2,267,404
Basophil proportion (14,096)	Basophil count (313)	1,647,073
Eosinophil proportion (19,760)	Eosinophil count (168)	1,579,993
Lymphocyte proportion (17,621)	Lymphocyte count (208)	1,590,215
Monocyte proportion (13,775)	Monocyte count (183)	1,601,135
Neutrophil proportion (17,622)	Neutrophil count (184)	1,586,187
Strata 3[c]:		
One indicates FBC but the other indicates anything other than FBC		638,650

[a] Where one code suggested a particular FBC parameter but the other suggested a different FBC parameter, with existing medical or entity codes for both parameters that could have been assigned

[b] Where the medical code suggested one of the six FBC parameters without an existing entity code

[c] Where one code suggested a particular FBC parameter but the other suggested the record was not FBC-related

commonly entity 189 (platelets), with 162,335 occurrences where the medical code indicated platelet distribution width (medical code 33,285), which describes the variation in the size of platelet cells. There were 83 different FBC medical codes assigned where the corresponding entity code was not FBC-related, commonly 64 (red blood cell distribution width), with 92,232 occurrences where the corresponding entity code was 289 (film report). From 14 FBC entity codes, we identified 609 parameters assigned medical code 0, which represents missing data (see Additional file 1: 2).

Availablility of mismatched parameters

In strata 1, the 44,349 mismatched pairs belonged to 44,221 FBC tests, with most tests having only one mismatched pair among them (99.7%, n = 44,098). Both parameters suggested by the medical and entity code were already available in that FBC for 96.7% (n = 42,891) of mismatched pairs, one of the two suggested parameters was available for 0.2% (n = 68), and neither were available for 3.1% (n = 1,390).

In strata 2, the 10,272,007 mismatched pairs were among 3,615,271 FBC tests. The majority of tests had only one mismatched pair (56.5%, n = 2,042,743), followed by five parameters (34.7%, n = 1,252,826). Among the mismatched pairs, the parameter suggested by the entity code was already available in that FBC test for 94.5% (n = 9,708,930) and neither parameter suggested by the medical and entity were available for 5.5% (n = 563,077). The parameter suggested by the medical code was not already present in any of the 3,615,271 FBC tests.

In strata 3, the 638,650 mismatched pairs were among 357,315 FBC tests. Most of these tests (64.3%, n = 229,836) had only one mismatched pair among them. Among the mismatched pairs, the parameter suggested by the medical code was already available

in the same FBC for 2.0% (n = 12,826) and the parameter suggested by the entity code already available for 47.2% (n = 301,292). For 50.8% (n = 324,532), neither parameter suggested by the medical or entity code was already available in the same FBC test.

Rectify mismatches

In strata 1, the entity code for 51 parameters and the medical code for four were chosen. For 34,059, it was not clear which of the two parameters the value and units represented. The remaining 10,235 mismatched pairs had no values entered and could not be assessed.

In strata 2, the entity code for 153,590 parameters and medical code for 9,471,831 were chosen. For 602,633 mismatched pairs, it was not clear which of the two parameters the value and unit represented. The remaining 43,953 had no values entered.

In strata 3, the entity code for 24,925 parameters and medical code for 113,065 were chosen. For 344,438, it was not clear what the value and unit represented. The remaining 156,222 mismatched pairs had no values entered.

Of the 10,955,006 mismatched pairs, the most common FBC parameter rectified was mean platelet volume (n = 2,041,360) and 1,191,540 could not be rectified (see Additional file 1: 3 for full details). The resulting dataset consisted of 227,515,914 consistently coded or rectified FBC parameters among 2,914,589 patients. Table 2 shows the total number of FBC parameters in the resulting dataset.

Standardising FBC units

In the resulting dataset, most units of measurement were unreliable, such as missing, partially entered, or clearly wrong. For example, red blood cell count values reported in seconds (Additional file 1: 4). Some parameters had values that seemed to be in the standardised unit but an alternative unit was recorded. Where there were extreme values, it was not possible to assess whether the unit was correct or if the value was associated with an alternative unit of measurement. The red blood cell count had the most variability in number of units (n = 89) and mean platelet volume had the least (n = 3) (Fig. 2). No apparent differences in the values and units of each FBC parameter over time were observed in the dataset.

Standardisation was not possible due to the high volume of inconsistent and incomplete units of measurement. Consequently, only parameters where the units were already entered as those we planned to standardise to were included in the dataset (Table 4 shows the final units of measure). The resulting dataset consisted of 81.3% (n = 185,982,456) parameters of the original 228,707,454 among 2,870,006 patients (Table 2).

To identify the extent of extreme or implausible values, summary statistics for deciles for each parameter were calculated, except for basophils, eosinophils, lymphocytes, monocytes, and neutrophils, where quartiles were derived. On the lower end, there were 171 parameters with negative values and 6,327,555 values entered as zero. On the higher end, for each parameter, the highest quantile showed a plausible median value and interquartile range, suggesting relatively few extreme values (see Additional file 1: 5).

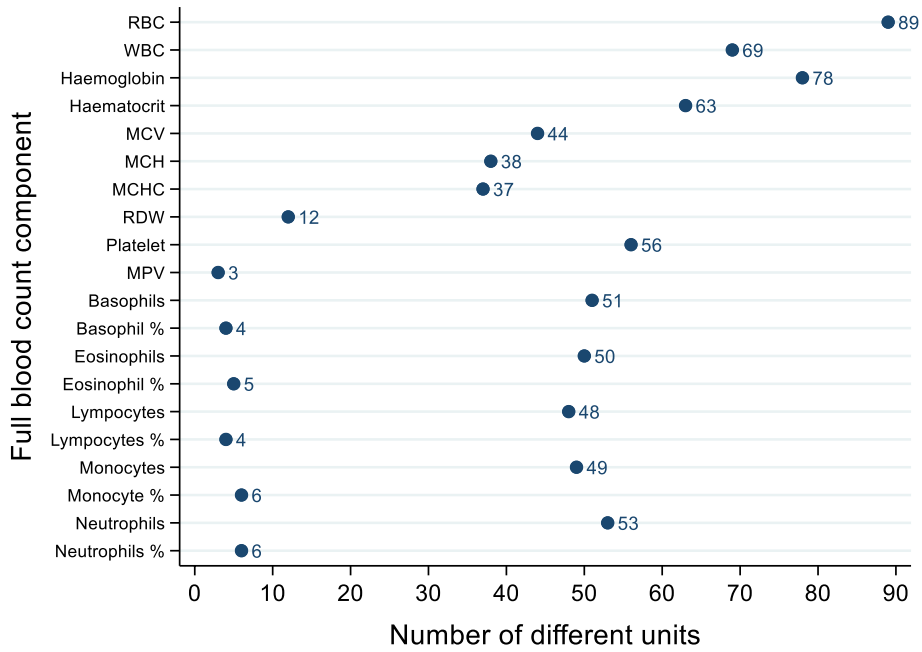

Fig. 2 Number of different units entered for each full blood count parameter (n = 769 in total).*RBC* red blood cell count, *WBC* white blood cell count, *MCV* mean corpuscular volume, *MCH* mean corpuscular haemoglobin, *MCHC* mean corpuscular haemoglobin concentration, RDW red blood cell distribution width, *MPV* mean platelet volume

Table 4 Descriptive statistics of each FBC parameter in the final derived dataset

FBC parameter	No. records	Mean	SD	Median	Min	Max	No. (%)[a] missing values
RBC (10^{12}/L)	14,989,333	4.3	5.3	4.4	0	8,577.0	1,547,684 (9.4%)
WBC (10^9/L)	13,391,464	7.6	46.6	6.9	0	123,000.0	3,145,553 (19.0%)
Haemoglobin (g/dL)	15,956,531	14.7	14.2	13.5	0	13,728.7	580,486 (3.5%)
Haematocrit (L/L)	14,929,673	30.6	49.0	38.7	0	121,881.6	1,607,344 (9.7%)
MCV (fL)	15,508,063	90.7	15.4	91.0	0	10,730.0	1,028,954 (6.2%)
MCH (fL)	14,978,027	45.4	214.4	30.3	0	33,378.5	1,558,990 (9.4%)
MCHC (g/dL)	14,758,966	318.3	973.6	33.5	0	352,127.7	1,778,051 (10.8%)
RDW (%)	312,052	14.0	1.7	13.7	0	43.8	16,224,965 (98.1%)
Platelets (10^9/L)	15,409,814	266	175.0	256.0	0	319,000.0	1,127,203 (6.8%)
MPV (fL)	1,964,340	9.6	7.3	9.4	0	726.0	14,572,677 (88.1%)
Basophil count (10^9/L)	11,532,657	0.1	2.4	0.1	0	6,500.0	5,004,360 (30.3%)
Basophil proportion (%)	11,577,467	0.9	1.7	0.8	0	2,694.1	4,959,550 (30.0%)
Eosinophil count (10^9/L)	11,609,391	0.2	0.7	0.2	0	1,218.0	4,927,626 (29.8%)
Eosinophil proportion (%)	11,619,040	3.1	3.4	2.6	0	5,869.8	4,917,977 (29.7%)
Lymphocyte (10^9/L)	12,202,709	2.1	78.8	1.9	0	275,000.0	4,334,308 (26.2%)
Lymphocyte proportion (%)	12,195,027	28.5	43.4	28.1	0	95,486.1	4,341,990 (26.3%)
Monocyte count (10^9/L)	12,068,645	0.6	0.6	0.5	0	1,218.0	4,468,372 (27.0%)
Monocyte proportion (%)	12,076,053	7.8	5.5	7.5	0	8,590.6	4,460,964 (27.0%)
Neutrophil count (10^9/L)	12,306,549	4.5	7.1	4.1	0	12,200.0	4,230,468 (25.6%)
Neutrophil proportion (%)	12,299,432	60.4	44.2	60.0	0	59,623.5	4,237,585 (25.6%)

FBC full blood count, *RBC* red blood cell count, *WBC* white blood cell count; *WBC* white blood cell count, *MCV* mean corpuscular volume, *MCH* mean corpuscular haemoglobin, *MCHC* mean corpuscular haemoglobin concentration, RDW red blood cell distribution width, *MPV* mean platelet volume

[a] Percentages are calculated out of the total number of tests (n = 16,537,017)

Derive missing FBC values

Initially, 332 parameters that had no test date were removed, as these parameters could not be matched to other parameters to derive missing values in the same FBC. A further 8,091,664 parameters were removed because multiple blood tests taken on the same day in a single patient could not be distinguished. The resulting dataset contained 177,890,460 parameters from 16,589,428 FBC tests among 2,856,400 patients.

Before applying the equations in Additional file 1: 1, there were approximately 11 parameters available on average among the 16,589,428 tests. After application, there were 16 parameters available on average. A parameter was considered unreliable if the derived value was negative and consequently considered all mathematically associated parameters to be unreliable, and deleted 23,541,367 parameters. At a blood test level, we deleted 52,411 FBC tests that had no available data for any individual parameter.

Describe FBC data

Of the original 2,914,589 patients, 2.0% (n = 58,569) of patients were excluded in total after all amendments. Of those excluded, the mean age at study entry was 42.2 years and 47.2% were male. The final dataset contained 241,685,233 parameters from 16,537,017 FBC tests from 2,856,020 patients. Mean age at study entry was 42.3 years and 44.5% were male in the final dataset.

There were approximately 15 parameters available per FBC on average. The FBC dataset contained more parameters than were originally available in the CPRD Test dataset. Table 2 shows the total number of each parameter in the final dataset. See Additional file 1: 5 for the number of tests for each number of parameters available. All nine measured parameters were all available for 67.1% (n = 11,102,834) and all 11 derived parameters for 0.3% (n = 48,046) of FBC tests. Only 0.3% (n = 47,999) of FBC tests had all 20 parameters available.

Summary statistics for each parameter are in Table 4. All parameters appeared to have extreme or implausible values, as indicated by their minimum and maximum values. Haemoglobin had the least amount of missing data, with 3.5% of tests having unknown values, and red blood cell distribution width had the most, missing for 98.1% of tests.

Discussion

Many research studies that use laboratory data from EHR datasets do not often report assessing the quality of the dataset analysed. We identified three studies that performed quality checks using FBC data in a recent review [18–21]. This study highlights the quality of data from EHR datasets should be assessed to ensure a fundamental understanding of the data and to derive a reliable dataset for analysis. This is further emphasised because the use of these databases for research was not the primary reason for their development. With application to FBC data from the CPRD Test dataset, approximately 5% of the data has mismatches in coding, with medical codes (translated from Read codes) and entity codes (from the EHR system) suggesting results from different tests or examinations performed in practice. The underlying procedure of assigning entity codes to patient records within the EHR are unknown and the process at haematology laboratories and primary care practices are automated, so it is unclear why there were some

mismatches. No other studies that explored the consistency of the two coding systems in CPRD were identified. Mismatched pairs did not belong to a particular practice or group of practices, with approximately 97% of practices in the CPRD dataset having at least one FBC test with an inconsistently coded parameter. Furthermore, no major differences were observed over time.

Quite often, one of two FBC parameters suggested by either code was already available for that FBC, but this did not necessarily mean that the other parameter is the incorrect one, as other mismatched pairs within that same FBC might could suggest that same parameter. Approximately 17% of the mismatched pairs were rectified based on the FBC value and corresponding unit, where plausible. Furthermore, standardising the blood values to conventional units was not possible because many were not appropriately recorded in CPRD, such as partially entered or did not appear to match the value. Some units were clearly wrong, such as blood values measured in seconds. However, as the majority of parameters were recorded in standard units, the proportion of parameters dropped was relatively low.

It is likely that many researchers are not aware of the mathematical relationship between parameters, with only one study identified using such equations to derive missing data [10]. A previous study has compared different approaches for missing data imputation of clinical laboratory measurements, including the full blood count, but do not discuss these mathematical relationships between parameters [25]. Using our approach to resolve missing data, we derived a dataset of FBC parameters that contained more data than originally available in the CPRD dataset of FBCs.

All 20 parameters are automatically derived from laboratory analysers in recent years, although this has not always been the case. This may explain why the original CPRD dataset had approximately 11 of 20 parameters available per FBC on average. Reasons for missing data are not recorded in CPRD, but one possible explanation is likely due to technology catching up to changes in practice as new parameters become available. After deriving missing data using known mathematical relationships between parameters, approximately 15 of 20 parameters were available per FBC on average but less than 1% of tests had all 20 parameters available. Of all 20 FBC parameters, missing data was most common for the red blood cell distribution width and mean platelet volume parameters, missing for approximately 98% and 88% of FBC tests, respectively. Historically, these parameters were derived by laboratory analysers along with the other 18 parameters but the output suppressed before the FBC report is sent to the GP practice. This was because the parameters were not considered helpful or meaningful, which was considered standard practice until recently. This could explain why many FBC tests in CPRD have missing data for these parameters.

Approximately 59,000 patients (2%) with FBCs were removed from the original CPRD Test dataset through our data quality check, resulting in a relatively large dataset of FBC results. Age and gender were the only demographic data available and were balanced between those included and those excluded, suggesting no differences in key patient characteristics. The final dataset is therefore considered representative of the overall sample.

A systematic review identified many barriers to perfoming quality checks. These include handling large amounts of unstructured data, problems with data extraction, and unfamiliarity with data quality assessment [15]. A second review highlighted a need for a generalised approach to assess EHR data quality [17]. Our methodological approach could help tackle these barriers to assessing data quality from EHRs and help researchs improve the quality of research findings.

Recommendations

Our methodological approach was applied using a dataset of FBCs from CPRD. However, the approach can form a basis and be adapted for researchers to assess the quality of other tests and examinations and from other datasets. To help researchers prepare their EHR datasets for analysis, we provide our Stata statistical programming (Additional file 1: 6) for the FBC data quality check for other researchers to make use of.

Often, EHR staff perform the data cut from the EHR and subsequently extract the relevant data items using the clinical codes used in the EHR. We recommended researchers extract the appropriate data items or have close involvement with EHR staff who extract the data to better identify the accuracy of the dataset and develop a fundamental understanding of the processes involved in preparing EHR datasets.

We recommend researchers use the mathematical equations to derive missing FBC data, thereby ensuring the relationship between parameters within a FBC holds. If subsequently there is still missing data, we suggest researchers use multiple imputation to impute the values of eight parameters: red blood cell count, haemoglobin, platelet count, basophil count, eosinophil count, lymphocyte count, monocyte count, and neutrophil count. This is because these parameters are measured from a blood sample and can be used to derive missing data for the other parameters using mathematical equations. Researchers should consider the need for inclusion of the red blood cell distribution width and mean platelet volume, for which missing data was common, because imputation may not be plausible and including FBCs with these parameters will drastically reduce the sample size.

One reason for limited data validation among research studies is that large datasets are computationally intensive. We recommend researchers invest in powerful laptops that are efficient for data processing and either internal or external hard drives for data storage. Furthermore, we recommend that researchers factor data quality checks into their study timelines, as the process can take many months but is crucial to ensure a reliable dataset for analysis.

Conclusion

Without performing data assessments, the opportunity to understand the dataset and assess its accuracy is often missed. We describe how there are a number of considerations when preparing EHR data and advise researchers to perform data quality checks to understand the extent of any issues, to derive a reliable dataset for analysis. Although routine datasets provide a large sample size for analysis, we emphasise that the reliability of the data should be prioritised.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40537-020-00375-w.

Additional file 1: **1.** Mathematical equations linking FBC components. **2.** Medical code of zero. **3.** Number of mismatched pairs that were rectified into a single parameter per FBC parameter. **4.** Number of components for each unit of measure per FBC parameter. **5.** Summary of quantiles and number of tests available per FBC parameter. **6.** Stata programming used to assess the quality of FBC data recorded in the Test dataset of the CPRD database

Abbreviations

CPRD: Clinical Practice Research Datalink; EHR: Electronic health records; FBC: Full blood count; GPSoC: GP Systems of Choice; ICD-10: International Classification of Diseases-10; SD: Standard deviation.

Acknowledgements

The authors would like to thank patient and public involvement representatives, Julian Ashton, Margaret Ogden, and Pete Wheatstone for their input in the interpretation of results and development of this manuscript.

Authors' contributions

PSV, TH, and JB developed the study. AF extracted the data from the CPRD database. PSV performed the quality check of the CPRD data. AF provided input regarding CPRD processes, MJ provided input regarding haematology laboratory processes, and TH provided input regarding clinical practice processes. PSV performed all analyses under the supervision of TH and JB. PSV drafted this manuscript. All authors read and approved the final manuscript.

Funding

PV is funded by a National Institute for Health Research (NIHR), Doctoral Research Fellow (DRF-2018-11-ST2-057) for this research project. JB is funded by the NIHR Oxford Biomedical Research Centre (BRC), Oxford University Hospitals NHS Foundation Trust. This report presents independent research and the views expressed are those of the authors and not necessarily those of the funders, NHS, or Department of Health and Social Care.

Availability of data and materials

The datasets used in this study are available from CPRD but restrictions apply [2].

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Centre for Statistics in Medicine, Botnar Research Centre, Nuffield Orthopaedic Centre, NDORMS, University of Oxford, Windmill Road, Oxford OX3 7LD, UK. [2] Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK. [3] BMS Haematology, John Radcliffe Hospital, Oxford University Hospitals, Oxford, UK.

Received: 3 July 2020 Accepted: 1 November 2020
Published online: 10 November 2020

References

1. GP Systems of Choice. 2019. <https://digital.nhs.uk/services/gp-systems-of-choice>. Accessed 11 Dec 2019.
2. Clinical Practice Research Datalink (CPRD). 2019. <https://www.cprd.com/>. Accessed 11 Dec 2019.
3. Clinical Practice Research Datalink (CPRD): Primary care data for public health research. 2019. <https://cprd.com/primary-care>. Accessed 11 Dec 2019.
4. Padmanabhan S. CPRD GOLD Data Specification. 2017 <https://cprdcw.cprd.com/_docs/CPRD_GOLD_Full_Data_Specification_v2.0.pdf>. Accessed 11 Dec 2019.
5. Watson J, Nicholson BD, Hamilton W, et al. Identifying clinical features in primary care electronic health record studies: methods for codelist development. BMJ Open. 2017;7(11):e019637. https://doi.org/10.1136/bmjopen-2017-019637.
6. Benson T. The history of the read codes: the inaugural James read memorial lecture. Inform Prim Care. 2011;19(3):173–82. https://doi.org/10.14236/jhi.v19i3.811.
7. World Health Organisation: Classification of disease. 2016. <https://icd.who.int/browse10/2016/en>. Accessed 11 Dec 2019.
8. Lab Tests Online: Full Blood Count (FBC). 2020. <https://labtestsonline.org.uk/tests/full-blood-count-fbc>. Accessed 30 May 2020.
9. Bailey SER, Ukoumunne OC, Shephard EA, Hamilton W. Clinical relevance of thrombocytosis in primary care: a prospective cohort study of cancer incidence using English electronic medical records and cancer registry data. Br J Gen Pract. 2017. https://doi.org/10.3399/bjgp17X691109.
10. Birks J, Bankhead C, Holt TA, Fuller A, et al. Evaluation of a prediction model for colorectal cancer: retrospective analysis of 2.5 million patient records. Cancer Med. 2017. https://doi.org/10.1002/cam4.1183.
11. Ankus E, Price SJ, Ukoumunne OC, Hamilton W, et al. Cancer incidence in patients with a high normal platelet count: a cohort study using primary care data. Fam Pract. 2018;35(6):671–5. https://doi.org/10.1093/fampra/cmy018.
12. Hippisley-Cox J, Coupland C. Identifying patients with suspected colorectal cancer in primary care: derivation and validation of an algorithm. Br J Gen Pract. 2012. https://doi.org/10.3399/bjgp12X616346.

13. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify men with suspected cancer in primary care: derivation and validation of an algorithm. Br J Gen Pract. 2013a. https://doi.org/10.3399/bjgp13X660724.
14. Hippisley-Cox J, Coupland C. Symptoms and risk factors to identify women with suspected cancer in primary care: derivation and validation of an algorithm. Br J Gen Pract. 2013b. https://doi.org/10.3399/bjgp13X660733.
15. Ni K, Chu H, Zeng L, Li N, et al. Barriers and facilitators to data quality of electronic health records used for clinical research in China: a qualitative study. BMJ Open. 2019;9:e029314. https://doi.org/10.1136/bmjopen-2019-029314.
16. Charnock V. Electronic healthcare records and data quality. Health Info Libr J. 2019;36:91–5. https://doi.org/10.1111/hir.12249.
17. Gray N, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144–51. https://doi.org/10.1136/amiajnl-2011-000681.
18. Virdee PS, Marian IR, Mansouri A, Elhussein L, et al. The full blood count blood test for colorectal cancer detection: a systematic review, meta-analysis, and critical appraisal. Cancers. 2020;12(2348):1–37. https://doi.org/10.3390/cancers12092348.
19. Boursi B, Mamtani R, Hwang WT, Haynes K, et al. A risk prediction model for sporadic crc based on routine lab results. Dig Dis Sci. 2016;61(7):2076–86. https://doi.org/10.1007/s10620-016-4081-x.
20. Firat F, Arslan AK, Colak C, Harputluoglu H. Estimation of risk factors associated with colorectal cancer: an application of knowledge discovery in databases. Kuwait J Sci. 2016;43(2):151–61.
21. Prizment AE, Anderson KE, Visvanathan K, Folsom AR. Association of inflammatory markers with colorectal cancer incidence in the atherosclerosis risk in communities study. Cancer Epidemiol Biomarkers Prev. 2011;20(2):297–307. https://doi.org/10.1158/1055-9965.EPI-10-1146.
22. CALIBER—Diseases of the blood. 2019. <https://www.caliberresearch.org/portal/chapter/5#Diseases%20of%20the%20blood>. Accessed 18 Dec 2019.
23. CllinicalCodes.org—Examining variations in prescribing safety in UK general practice: a cross-sectional study using the Clinical Practice Research Datalink. 2019. <https://clinicalcodes.rss.mhs.man.ac.uk/medcodes/article/25/codelist/res25-m3_lft_fbc/>. Accessed 18 Dec 2019.
24. Iwagami M, Caplin B, Smeeth L, Tomlinson LA, et al. Clinical codelist - read codes for severe mental illness. London: London School of Hygiene & Tropical Medicine; 2018. (**10.17037/DATA.00000868**).
25. Abdala OT, Saeed M. Estimation of missing values in clinical laboratory measurements of ICU patients using a weighted K-nearest neighbors algorithm. Comput Cardiol. 2004;31:693–6.

Publisher's Note