Pharmacoepidemiology and Drug Safety

WILEY

ORIGINAL ARTICLE OPEN ACCESS

A New Validated Approach for Identifying Childhood Immunizations in Electronic Health Records in the United Kingdom

Anne M. Suffel[1,2] | Jemma L. Walker[1,2,3] | Colin Campbell[3] | Helena Carreira[1] | Charlotte Warren-Gash[1] | Helen I. McDonald[1,2,4]

[1]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK | [2]NIHR Health Protection Research Unit in Vaccines and Immunisation at London School of Hygiene and Tropical Medicine, London, UK | [3]UK Health Security Agency, London, UK | [4]Department of Life Science, University of Bath, Bath, UK

Correspondence: Anne M. Suffel (anne.suffel@lshtm.ac.uk)

Received: 17 November 2023 | **Revised:** 13 May 2024 | **Accepted:** 23 May 2024

Funding: This study is funded by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Vaccines and Immunisation (NIHR200929), a partnership between the NIHR and Wellcome. The views expressed are those of the author(s) and not necessarily those of the NIHR, UK Health Security Agency, or the Department of Health and Social Care. C.W.-G. is supported by a Wellcome Career Development Award (225868/Z/22/Z).

Keywords: childhood immunizations | electronic health records | real-world evidence | vaccine

ABSTRACT

Background: Routinely collected electronic health records (EHR) offer a valuable opportunity to carry out research on immunization uptake, effectiveness, and safety, using large and representative samples of the population. In contrast to other drugs, vaccines do not require electronic prescription in many settings, which may lead to ambiguous coding of vaccination status and timing.

Methodology: We propose a comprehensive algorithm to identifying childhood immunizations in routinely collected EHR. In order to deal with ambiguous coding, over-recording, and backdating in EHR, we suggest an approach combining a wide range of medical codes in combination to identify vaccination events and using appropriate wash-out periods and quality checks. We illustrate this approach on a cohort of children born between 2006 and 2014 followed up to the age of five in the Clinical Practice Research Datalink (CPRD) Aurum, a UK primary care dataset of EHR, and validate the results against national estimates of vaccine coverage by NHS Digital and Public Health England.

Results: Our algorithm reproduced estimates of vaccination coverage, which are comparable to official national estimates and allows to approximate the age at vaccination. Electronic prescription data only do not cover vaccination events sufficiently.

Conclusion: Our new proposed method could be used to provide a more accurate estimation of vaccination coverage and timing of vaccination for researchers and policymakers using EHR. As with all observational research using real-world data, it is important that researchers understand the context of the used dataset used and the clinical practice of recording.

1 | Introduction

Vaccinations prevent over 3.5–5 million deaths every year and hence, are one of the most successful public health interventions [1]. Nevertheless, the global uptake of childhood immunizations has plateaued in the last decade and only 11 countries met a coverage of at least 90% for all recommended vaccines in 2019 [2]. Additionally, many vaccination services were interrupted due to

Charlotte Warren-Gash and Helen I. McDonald share joint senior authorship.

Summary

- The recording of vaccines in electronic health records is strongly influenced by setting where the vaccine is administered and data flow between different care settings.
- Prescription data only do not cover vaccination sufficiently.
- A combination of a wide range of vaccine-related medical codes can help to deal with ambiguous coding.
- Minimum age requirements and wash-out periods between vaccine doses are useful to differentiate several doses from double-recording.
- Our comprehensive algorithm for cleaning vaccine-related electronic health records achieved comparable results to national coverage estimates.

the COVID-19 pandemic with up to 25 million children worldwide who missed out on vaccination [1].

Population-based observational studies of vaccination are important to maintain and assure immunization programs, including understanding and improving vaccine uptake. Furthermore, high-quality studies of real-world vaccine effectiveness and safety are essential to protect health and build public confidence [3]. As the determinants of vaccination are complex and very context-specific [4], detailed population-wide data on vaccine uptake can help to design multicomponent interventions that target disadvantaged populations directly [5].

As there are a variety of different settings where childhood immunizations can be administered, this implies a careful choice of the data source for a vaccine study. In the United Kingdom, routine childhood vaccinations up to the age of 5 are largely delivered and recorded in primary care [6]. However, vaccines can be delivered in a variety of different settings: for children, this includes schools for children aged 5 years and over, secondary care for high-risk groups who are recommended additional vaccines, and pharmacies. Other vaccines, such as travel-related vaccines or the chickenpox vaccine, are mainly given privately in pharmacies or private clinics.

In theory, UK primary care data should include a record of all vaccinations, wherever received. However, the completeness of recording in primary care depends on the methods of data collection and transfer from other vaccinating settings, which are variable in schools [6] and antenatal clinics [7], resulting in incomplete or delayed immunization records. Vaccinations given in other settings, such as pharmacies, might not be well recorded at all.

Electronic health records (EHR) are routinely collected from different health care setting which can cover a large part of the population [8, 9]. As vaccines are delivered as part of routine health care, EHR have increasingly gaining importance in vaccine research [10–12]. EHR can help to explore inequalities in uptake [13], investigate the impact of changes in vaccination schedules [14, 15], and assess real-world vaccination effectiveness and safety [16, 17].

Understanding the context of data collection and curation is a general challenge for high-quality EHR research as these routine datasets were not collected for research purposes but clinical care and administrative processes [18]. For example, in the Clinical Practice Research Datalink (CPRD) Aurum, an electronic health record dataset for primary care data in the United Kingdom, several codes can be used to encode the same vaccine and can be saved as prescription data or observation data [9].

This might lead to ambiguous or conflicting coding of immunizations and researchers might miss important codes by focusing on prescription data only.

In this article, we present a comprehensive algorithm to identify vaccination events using an example of a cohort of children born between 2006 and 2014 in England from CPRD Aurum, validate our results by comparison to the national estimates of the national health authorities [19] and make recommendations for vaccine researchers and policymakers.

2 | Methods

2.1 | The Electronic Health Record Dataset

Primary care datasets such as CPRD Aurum provide rich, longitudinal patient-level information from a primary care setting on symptoms and diagnoses, clinical tests and results, immunizations, prescriptions, and referrals to other services [9]. CPRD Aurum contains anonymized data originally collected from primary care practices in England and Northern Ireland using the EMIS Web electronic patient record system software to manage patient care. In 2022, CPRD Aurum contained data from around 41 million patients and was broadly representative in geographical spread, age, sex, and ethnicity [9, 20]. Data on clinical diagnoses, symptoms, clinical tests, and referrals are collected in an observation table using SNOMED CT, Read Version 2, and local EMIS Web codes, whereas data on drug prescriptions and devices are collected in a separate drug issue table coded using Dictionary of Medicines and Devices (DM + D) [9]. Coded records for clinical diagnoses in CPRD usually show a high validity in CPRD GOLD but less information is available for CPRD Aurum [21].

The data from practices can be linked to other datasets including secondary care data, death registries, and either patients' or local practices' area-level measures of relative deprivation [9, 22].

2.2 | The Study Population and Vaccines of Interest for Validating the Algorithm

Childhood vaccine schedules have been changed and adapted many times over the recent years and will be subject to ongoing changes [23]. In order to compare the results of our algorithm against national estimates for vaccine coverage, we chose a time period during which different routine vaccines were consistently given and did not underly changes of the vaccination schedule. Additionally, we avoided follow-up during the COVID-19 as the pandemic was likely to disrupt both data collection and service delivery. Hence, we focused on the antigens

tetanus, diphtheria, pertussis, pneumococcus, measles mumps, and rubella as those have been consistently recommended in England at the same ages between 2006 and 2019 [23].

To validate our algorithm, we used anonymized data from the CPRD Aurum of all children registered with a GP practice contributing to CPRD Aurum who were born between 2006 and 2014. Each child was followed up until they either changed GP practice, died or turned 5 years old. This dataset provided clinical records and prescription data for each child from registration until end of follow up. To preserve confidentiality, CPRD provides only month and year of birth for children rather than an exact date of birth. Therefore, to better estimate the timing of each vaccination, their date of birth was set to the 15th of their month of birth. After conducting quality checks which entailed using an acceptability flag provided by CPRD, we removed children who had a record of vaccination within 2 weeks around their estimated date of birth as we interpreted this as indicator of unreliable vaccination recording. The final study population included 1 735 692 individuals.

A summary of vaccination schedule of consistently administered vaccines during the study period and the plausible age range for each appointment due to unknown date of birth can be found in Table 1.

2.3 | The Vaccination Event Algorithm

2.3.1 | Types of Codes and Data Used for Vaccinations

Vaccination data in CPRD can be either stored in the prescription table or in the observation table using different coding systems. If a vaccine is administered in the same practice where the patient is registered, no electronic prescription of the vaccine is necessary for dispensation using Patient Group Directions (PGD) and Patient Specific Directions (PSD) [24, 25]. Hence, a combination of prescription data (later called vaccine products) and clinical-coded data in the observation table is necessary to fully capture all administered vaccines. These can be linked via a pseudonymized patient identifier.

2.3.2 | Code List Considerations

2.3.2.1 | Creating of a Vaccine Code List. For each vaccine of interest, code lists have to be created based on search terms for antigens of interest, the infection vaccinated against, and brand names of vaccine products. These code lists will then be used to extract the vaccine-related information from the medical record. A more detailed general instruction how to create a code list in general can be found in our methodological article [26].

Considerations regarding what kind of codes to be included into the final code lists are explained in more detail in the following sections.

All search terms and created code lists for this study can be found online [27].

2.3.2.2 | Handling Combined Vaccinations. Many vaccines contain several antigens, for example, the 6-in-1, or the MMR vaccine [23]. In order to explore the necessity of creating code lists based on all combined antigens (e.g., measles + mumps + rubella vaccine) or whether using an example antigen is sufficient (e.g., measles vaccine), we created two types of code list, applied the algorithm below and compared the number of vaccination events identified by different code list. This was done for the measles antigen versus measles + mumps + rubella vaccine and for the pertussis antigen versus diphtheria + tetanus + pertussis vaccine.

2.3.2.3 | Using General Vaccination Codes. Furthermore, there are very general vaccination codes such as "childhood

TABLE 1 | Childhood vaccines examined in the study and the recommended age allowing for 16 days imprecision around the estimated date of birth.

Appointment	Vaccines[a]	Age range for the scheduled appointment (days)
Birth	—	−16 to 16
First appointment (56 days)	DTP	40–72
	PCV	
Second appointment (84 days)	DTP	68–100
Third appointment (112 days)	DTP	98–128
	PCV	
Fourth appointment (365 days)	MMR	349–381
	PCV	
Fifth appointment (1215 days)—booster doses	MMR	1199–1259
	DTP	

Abbreviations: DTP, diphtheria, tetanus, pertussis vaccine; MMR, measles, mumps, rubella vaccine; PCV, pneumococcal conjugate vaccine.
[a]These are vaccines selected to mark the appointment but not the only vaccines usually given at these appointments. As the date of birth had to be estimated, there is 16-day range of uncertainty around the true age into either direction.

immunization." In order to test whether specific vaccination events were replaced by very general codes for childhood immunizations in clinical practice, we compared the performance of two different code lists, one containing codes for the DTP vaccine only and a code list combining general vaccination terms plus DTP-specific codes. Consequentially, we performed our vaccination algorithm with both types of code lists and compared the number of vaccination events detected.

2.3.3 | Dealing With Conflicting or Ambiguous Types of Vaccination Records

Vaccine-related codes in primary care cannot only refer to administering a vaccine but also to discussions about vaccination, declined vaccination, adverse events, or recall systems. Consequentially, we investigated the impact of different categories of vaccine codes on defining vaccination events.

First, we explored the commonly used SNOMED codes in relation to the MMR vaccine and looked for recurring patterns of code. Consequentially, we decided to apply the following four categories of codes: codes for an actively given vaccine (indicated by words used such as "vaccine given" or "vaccine administered"), neutral vaccination codes (e.g., "MMR vaccine"), product codes (from prescription records) indicating a specific pharmaceutical product used, and codes for declining a vaccine. We counted codes just naming the type of vaccine (e.g., "measles vaccine") as neutral as they can be used together with other codes indicating a discussion or invitation to the vaccine. Table 2 presents an example for each of these categories for the MMR vaccine.

We chose not to include any codes about invitations to immunizations or recall as they add no information on whether a vaccine was given. It can be debated whether codes for adverse events after vaccination should be included as they clearly indicate that a vaccine was given at some point. However, we decided against including them into our algorithm as they these codes were rarely used (e.g., only 103 adverse events in 13 years of administering the MMR vaccine) and they added little information on the vaccine timing for the study.

Records which are neutral as to whether a vaccine was administered should be interpreted differently if they are recorded together with a prescription or a declined vaccination record. To determine if a vaccine had been delivered using the combined records on any given day the following algorithm was applied: for each patient the number of administered, neutral, product, and declined codes were summarized by date of the observation. Different combinations of codes led to interpretation of a child being vaccinated, a declined vaccination or conflicting codes. A neutral code might combined with a prescription or vaccine administered code to confirm vaccine delivery that day, or combined with a vaccine declined code to be interpreted as a vaccine not given that day (for all combinations of code categories, see Table S1). A combination of administered code and declined code was interpreted as conflict and the event excluded from the

TABLE 2 | Examples of different codes for the MMR vaccine falling under different categories and the proportion of codes falling under each category by vaccine type.

	Administered vaccine code	Neutral vaccine code	Vaccine product code	Vaccine declined code
Example codes	• Measles vaccination given • Administration of measles + mumps + rubella live vaccine • Vaccination for mumps given	• Measles vaccine • Measles mumps and rubella vaccination—first dose • Measles virus live attenuated	• Ervevax vaccine powder and solvent for solution for injection 0.5 mL vials (GlaxoSmithKline UK Ltd) 1 vial • Mumpsvax Injection • Priorix vaccine powder and solvent for solution for injection 0.5 mL vials (GlaxoSmithKline UK Ltd)	• Did not attend DTaP, polio, and MMR booster • No consent—measles immunization • MMR declined
Number of code type recorded by vaccine				
Number of codes extracted for pertussis vaccine	2 812 088 (44.42%)	3 451 632 (54.52%)	57 228 (0.90%)	10 408 (0.16%)
Number of codes extracted for pneumococcal disease vaccine	74 778 (1.56%)	4 669 056 (97.4%)	41 008 (0.85%)	6347 (0.13%)
Number of codes extracted for measles vaccine	1 430 800 (50.49%)	1 352 896 (47.47%)	26 462 (0.93%)	23 511 (0.82%)

analysis. A neutral, product, or administered code alone were interpreted as vaccination event.

2.3.4 | Determining the Number of Vaccine Doses and Timing of Administration

Most vaccines which are part of the NHS childhood immunization schedule consist of several doses, which should be given within a certain time period and usually require a minimum gap between doses to ensure maximum efficacy [23]. In order to evaluate the success of vaccination program, it is important to determine how many doses have been administered and whether they have been administered within the correct time interval. In real-world data, this is challenging as vaccine dose number is not typically recorded. Furthermore, determining the correct timing of a vaccine with respect to the child's age is important as some vaccines are not effective when given too early [28] or lead to a higher risk of transmission when delayed [29] or adverse effects from the vaccine [30]. For UK childhood vaccinations, the minimum intervals between vaccinations may be as little 4 weeks [23, 31].

The first difficulty is that it might not be possible to detect the precise age at vaccination in anonymized datasets as the exact day of birth may be not provided, for example being suppressed in CPRD Aurum to protect patient confidentiality [9]. We dealt with this issue by setting every child's day of birth to the 15th of the respective month and then allowing a range of 16 days uncertainty for every interval.

The second difficulty is that after identifying potential vaccination events combining different categories of vaccine codes, it is essential to differentiate separate vaccination events from multiple records of the same event. Particularly by using vaccination codes which we classified as "neutral," there is still a possibility that one of the codes referred only to the discussion of the vaccination or any other more administrative issue or side effects related to the vaccine. These records may result in apparent second doses.

For dealing with challenges of multiple reporting of the same event in EHR and short-time intervals between vaccines, we developed an algorithm which applied a combination of minimum age for each dose, and a minimum time interval between different vaccination doses, to define the timing of vaccine doses.

Figure 1 summarized the overall steps of the algorithm and Figure 2 illustrates the concept of the applied minimum age and age gap requirements. More detailed descriptions of each algorithm step with an example and code can be found in the Supporting Information S1.

Step 2 of the algorithm is illustrated in Figure S1. It shows the probability density of recording a vaccination event by different ages on a population level. It illustrates that especially early doses have to be differentiated carefully as there might be overlap of vaccine doses if definitions are based on age alone. However, all different doses show distinct peak ages of uptake.

Figure S2 shows how many vaccination records were removed at every stage of the algorithm.

Table S7 gives a summary of recommended minimum age and age gaps for different vaccines in our example. After applying this algorithm, there were still some individuals who had more than the four recommended pertussis vaccine doses recorded. A total of 3654 children had a total of 5 vaccine doses recorded, 169 children 6 doses and 16 children had a total of 7 doses recorded. All fifth doses and higher were dropped from the dataset. A potential explanation for the recording of more doses could be double recording of the same dose, maybe additional doses, which were given after exposure to one of the antigens in the combined vaccine product or a potential misclassification of the event itself, that is, only discussion regarding a vaccination instead of actually administering the vaccine.

3 | Results

3.1 | Cohort Summary

The final study cohort consisted of 1 735 692 children from 1474 practices in England, with median follow-up 1755 days (IQR 830–1800). This included 573 015 children followed up from 40 days after birth until the age of 5. A summary table of the demographics of the included children can be found in Table S1.

3.2 | Use of Different Code Lists

Using a single antigen-based code lists in comparison to a code list containing codes for all antigens of a combined vaccine detected more than a quarter fewer vaccination events in the dataset. However, there was almost no change in the number of children with vaccination events identified as many children have multiple records of vaccination on the same day (see Table 3).

However, this had almost no consequence on a population level. Table 4 illustrates how age of vaccination coverage and age at vaccination differed by code list used.

3.3 | Ambiguous Coding

In our cohort, the profile of different code categories used differed by antigen, for example with codes which were "neutral" as to whether the vaccine was given comprising 98% of records of pneumococcal antigen, and approximately half the records for pertussis and measles antigen (Table 2). Less than 0.2% of all the extracted codes for different vaccines were codes from the prescription table.

After combining the different code categories, very few conflicts were observed which means a vaccine declination code recorded on the same day with either a prescription or a vaccine administration code (e.g., 327 for the MMR vaccine [<0.01%]). We did not find any instances of three different code combinations on the same day.

All observed code category combinations are summarized in Table S5.

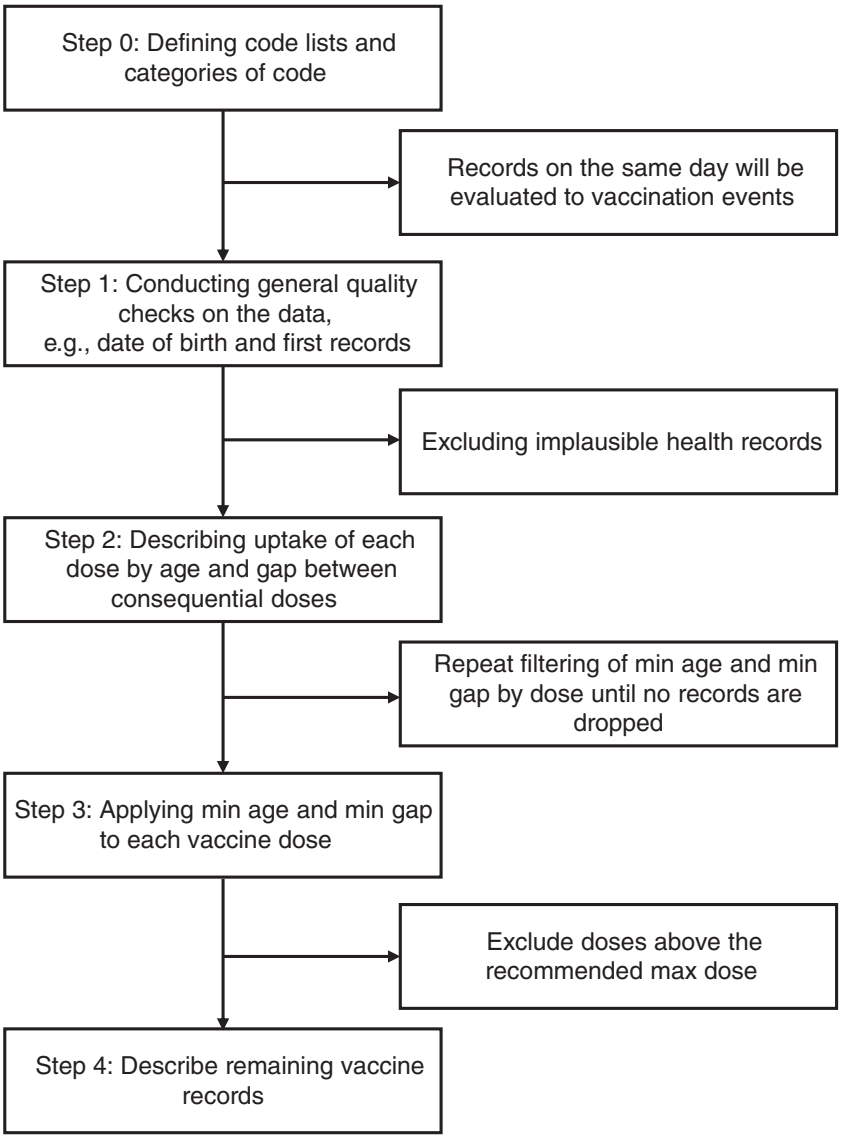

FIGURE 1 | Overview of the different steps in the data cleaning algorithm. Step 0 shows the interpretation of ambiguous codes as described above (Step 0). A vaccination record was interpreted as vaccine dose if given after a predefined minimum age and applying a minimum time interval between two doses.

3.4 | Performance of The Algorithm in Comparison to National Estimates

Table 5 summarizes the vaccination coverage at the age of 2 for the PCV vaccine and at the age of 5 for the MMR and DTP vaccine next to the national estimates of England. Our algorithm led to a very similar or even marginally higher vaccine uptake in the study population in comparison to national estimates. This finding was consistent across all types of vaccine.

A summary of age at vaccine receipt and gaps between different vaccines doses can be found in Table S6 using the pertussis vaccine as an example.

4 | Discussion/Conclusion

Overall, in this article, we demonstrated a comprehensive approach to identifying childhood immunizations in EHR which achieved comparable results to national estimates in England. Furthermore, we consider the complexity of vaccines and vaccine recording in EHR when designing vaccine studies using these datasets. On the example of a cohort from CPRD Aurum in England, we showed how different methods of identifying vaccines in primary care EHRs affect the estimates of vaccine uptake.

In-depth knowledge of the vaccine delivery systems is required to ensure that the right data source is chosen and to consider potential biases in data collection and reporting. Vaccine schedules can change over time and special situations such as outbreaks or exposure to different pathogens might impact when a vaccine was given and have to be taken into account. This approach must be adapted to the vaccine of interest and country of the study accordingly.

When creating a code list in order to identify vaccination events in EHR, a broader approach should be taken instead of focusing on vaccine prescriptions only. We showed that using an

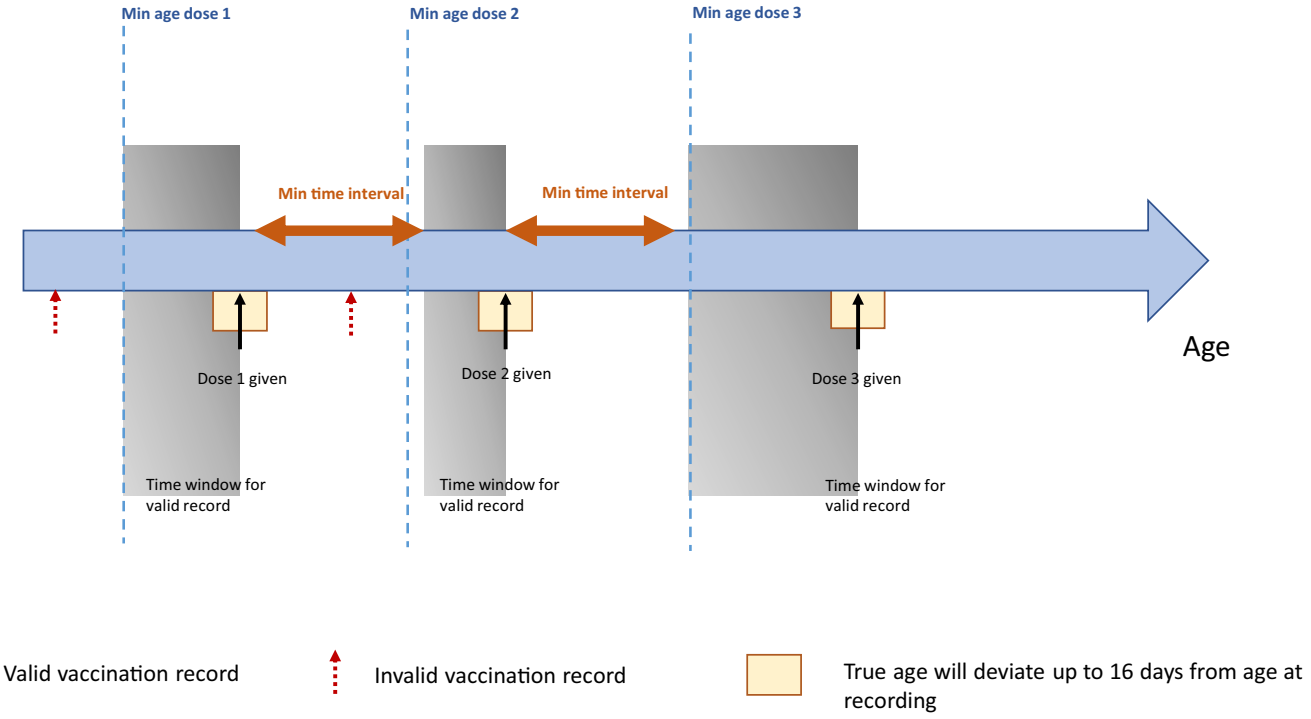

FIGURE 2 | Illustration of which vaccination records would be interpreted as valid vaccine doses after applying the vaccination algorithm.

TABLE 3 | Number of vaccination events and patients with vaccination detected by different types of vaccine code list.

	Number of vaccine-related events in dataset	Number of patients with any vaccine-related record
Code list containing all codes for measles, mumps, and rubella antigen containing vaccines	3 955 205 (100%)	1 531 523 (100%)
Code list containing codes for measles antigen containing vaccine only	2 894 040 (−26.82%)	1 531 396 (−0.00%)
Code list containing all codes for diphtheria, tetanus, and pertussis antigen containing vaccines	9 058 339 (100%)	1 702 697 (100%)
Code list containing codes for pertussis antigen containing vaccine only	6 389 450 (−29.46%)	1 700 423 (−0.00%)

TABLE 4 | Vaccine uptake and coverage for the MMR vaccine described using two different types of code lists to identify MMR vaccine-related health records.

Vaccine type	Dose	Mean age of receipt in days	Median age of receipt in days	IQR in days	Coverage at the age of 2 (in %)	Coverage at the age of 5 (in %)
MMR (measles antigen based)	1	455.50	409	387–446	93.86	97.55
	2	1273.00	1269	1236–1321	2.25	90.23
MMR (all antigen based)	1	455.40	409	387–446	93.87	97.55
	2	1272.00	1269	1236–1320	2.34	90.28

Abbreviations: IQR, interquartile range; MMR, measles, mumps, rubella vaccine.

algorithm based on a combination of prescription codes, codes indicating a vaccine refusal, neutral vaccine codes, and codes for given vaccines can show similar results to national vaccine coverage estimates if wash-out periods and minimum age gaps between the vaccine codes are applied. Nevertheless, there is some diversity in clinical coding which has to be considered when defining these categories and might not generally apply to all primary care practices even when using the same coding software [32]. Additionally, for applying this algorithm to other vaccines, the context where a vaccine is given and how

TABLE 5 | Overall vaccine coverage by calendar year from the vaccination algorithm in comparison to the NHS digital estimates of the same year in England.

Year	DTP vaccine	National estimates DTP[a]	Pneumococcal vaccine[a]	National estimates Pneumococcal[a]	MMR vaccine coverage	National estimates for MMR[a]
2009–10	—	84.8	87.09	87.6	—	82.7
2010–11	—	85.9	88.65	89.3	—	84.2
2011–12	87.03	87.4	90.45	91.5	90.04	86.0
2012–13	87.99	88.9	92.47	92.5	90.78	87.7
2013–14	89.00	88.7	92.99	92.4	91.56	88.3
2014–15	89.42	88.5	93.24	92.2	91.23	88.6
2015–16	89.46	86.3	92.92	91.5	90.50	88.2
2016–17	89.21	86.2	92.90	91.5	89.91	87.6
2017–18	88.91	85.6		91.0	89.37	87.2
2018–19	88.82	84.8		90.2	89.20	86.4

Abbreviations: DTP, diphtheria, tetanus, pertussis vaccine; MMR, measles, mumps, rubella vaccine.
[a]Coverage for the pneumococcal vaccine was measured for three doses at the age of 2. Coverage for DTP and MMR refers to the completion of the preschool booster (fourth and second dose, respectively) at the age of 5.

this translates to appropriate data source. For example, the BCG vaccine for tuberculosis used to be given in hospitals only and hence, would have not been represented appropriately in primary care datasets [33].

This extensive data cleaning is necessary to avoid double counting of vaccine events or lagged recording in EHR and to interpret records, which are unclear as to whether the vaccine was administered. Considering more general vaccine for childhood immunization may help to identify declined vaccinations.

We consider that the slightly higher uptake of some vaccines in our study compared with national estimates might be due to selection bias in our study population which requires a constant registration with the same GP practice during follow-up, and hence, only captures children who did not move house within the study period. Furthermore, differences in the national estimates could be due to their broader definition of acceptable codes and different deduplication methods [34].

4.1 | Strengths and Limitations

Capturing vaccine status accurately is essential for high-quality studies investigating vaccine effectiveness and safety. We provided a practical guide to help researchers making decisions on how to identify vaccination events, and data management to support high-quality studies using EHR, supported by data from a worked example. We provided recommendations for each stage of the study design and provided step-by-step instructions for the data cleaning. In our example, we managed to replicate national estimates for vaccination coverage using our algorithm with only minor deviations.

Key limitations include the focus of this study on the UK health care system. Immunization schedules, delivery settings and coding are likely to differ between countries. However, our algorithm framework offers an approach which can be adapted to the setting. Depending on the dataset used, some information might not be accessible. In our example, there is 16-day uncertainty around the actual age at vaccination due to suppression of date of birth in CPRD. If a vaccination event was recorded more than once, it is also not possible to determine which of the events was the true vaccination event and which one might have been a consultation only. Our suggested approach using primary care records does not account for vaccines, which were given outside of a primary care setting, for example, vaccines in private clinics, BCG vaccines in hospitals, or emergency tetanus vaccinations after injury or potential exposure.

4.2 | Key Recommendations for Researchers

Overall, we recommend that researchers explore the setting and the delivery pathway for their vaccines of interest first at the planning stage of the study and before deciding on a dataset. After choosing an appropriate dataset, researchers in a UK setting should not use prescription data only but a combination of different vaccine-related codes provided in the EHR. Antigens

which are usually administered together could serve as proxy for each other if special indications for certain antigens are considered. Vaccine schedules and recommendations as well as delivery pathways undergo frequent changes which require up-to-date information for the respective study period and setting.

4.3 | Policy Recommendations

We recommend for policymakers, a more centralized system for collecting vaccination data in order to bridge the gap between vaccines administered in different setting of care. Linkage with medical records would help to improve vaccine surveillance and signal detection of adverse events in children and to explore vaccine coverage in groups of special need and high-risk groups. There have been attempts to bring more vaccine data and general health data together. For children in England, the Child Health Information Services (CHIS) have been set up to collect local clinical care records of all children in an area including immunizations but do not include wider clinical data which would be necessary for safety studies or to study-specific risk groups, and cannot link children who move between different regions [35]. Even less data are available for people over 19 years. Other vaccine-specific data collection services such as The National Immunization Management Service (NIMS) have initially focused on collecting data on COVID-19 and the influenza vaccine only [36].

The introduction of childhood immunizations into the Quality and Outcomes Framework (QOF) in 2021 might help to improve the data quality on conflicting or ambiguous coding [37].

Both in the United Kingdom and elsewhere, adopting centralized data collection systems for vaccinations linked to other sources of health care data will help to strengthen resources available for vaccine research studies.

4.4 | Plain Language Summary

Patients can receive vaccines in various places such as at their GP practice, hospitals, pharmacies, or schools. This influences how and where information on these vaccinations is stored in electronic health records. These codes for vaccinations can sometimes be ambiguous, and sometimes codes from one care setting are transferred to the GP practice with some delay. This study presents stepwise instructions on how researchers should study electronic vaccination data, combining different ways of how these vaccinations are coded and checking the recording dates of the vaccination against the vaccination schedule recommended by the health care provider.

We applied our approach to a cohort of children in England born between 2006 and 2014 and followed them up until their fifth birthday. The results from our algorithm were comparable to the national vaccination coverage estimates published by the National Health Service.

A validated method to gain information on vaccinations from GP practices is essential to study whether vaccines are safe after they are given to the wider population and to ensure that different groups of the population equally manage to get access to the vaccines.

Author Contributions

A.M.S., J.L.W., H.I.M., and C.W.-G. conceptualized the study and contributed to the study design. A.M.S. analyzed and interpreted the data and drafted the first draft of the article. C.C. added context from a provider perspective. The article was critically revised by all authors. All authors approved the submission of the article.

Acknowledgments

This work uses data provided by patients and collected by the NHS as part of their care and support. Using patient data are vital to improve health and care for everyone. There is huge potential to make better use of information from people's patient records, to understand more about disease develop new treatments, monitor safety, and plan NHS services. Patient data should be kept safe and secure, to protect everyone's privacy, and it is important that there are safeguards to make sure that it is stored and used responsibly. Everyone should be able to find out about how patient data are used. We would like to thank Jennie Johnson and the PRIMIS team for their insights on clinical coding practices.

Ethics Statement

We received data governance approval from CPRD (protocol number 22_001706) and ethical approval from the London School of Hygiene and Tropical Medicine's research ethics committee (reference number 27651). Data are collected for routine clinical purposes rather than research so individual consent is not required to collect data. Consent to contribute de-identified data to CPRD for research purposes is given at practice level. Individual patients have the right to opt out of sharing their de-identified data for research. For more information see the following link: Safeguarding patient data | CPRD. All methods were carried out in accordance with relevant guidelines and regulations by CPRD, the Medicines and Healthcare products Regulatory Agency (MHRA), and the ethics committee at LSHTM.

Consent

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The study uses data from the Clinical Practice Research Datalink (CPRD). CPRD does not allow the sharing of patient-level data. The data specification for the CPRD dataset is available at: https://cprd.com/cprd-aurum-may-2022-dataset. The code lists can be found at: https://github.com/Eyedeet/vaccine_methods_ehr_public/tree/main/codelists.

References

1. World Health Organization, "Immunization Coverage," accessed May 9, 2022, https://www.who.int/news-room/fact-sheets/detail/immunization-coverage.

2. N. C. Galles, P. Y. Liu, R. L. Updike, et al., "Measuring Routine Childhood Vaccination Coverage in 204 Countries and Territories, 1980–2019: A Systematic Analysis for the Global Burden of Disease Study 2020, Release 1," *Lancet* 398, no. 10299 (2021): 503–521, https://doi.org/10.1016/S0140-6736(21)00984-3.

3. A. de Figueiredo, C. Simas, E. Karafillakis, P. Paterson, and H. J. Larson, "Mapping Global Trends in Vaccine Confidence and Investigating

Barriers to Vaccine Uptake: A Large-Scale Retrospective Temporal Modelling Study," *Lancet* 396, no. 10255 (2020): 898–908, https://doi.org/10.1016/S0140-6736(20)31558-0.

4. H. J. Larson, C. Jarrett, E. Eckersberger, D. M. D. Smith, and P. Paterson, "Understanding Vaccine Hesitancy Around Vaccines and Vaccination From a Global Perspective: A Systematic Review of Published Literature, 2007–2012," *Vaccine* 32, no. 19 (2014): 2150–2159, https://doi.org/10.1016/j.vaccine.2014.01.081.

5. T. Crocker-Buque, M. Edelstein, and S. Mounier-Jack, "Interventions to Reduce Inequalities in Vaccine Uptake in Children and Adolescents Aged <19 Years: A Systematic Review," *Journal of Epidemiology and Community Health* 71, no. 1 (2017): 87–97, https://doi.org/10.1136/jech-2016-207572.

6. K. Tiley, E. Tessier, J. M. White, et al., "School-Based Vaccination Programmes: An Evaluation of School Immunisation Delivery Models in England in 2015/16," *Vaccine* 38, no. 15 (2020): 3149–3156, https://doi.org/10.1016/j.vaccine.2020.01.031.

7. A. Llamas, G. Amirthalingam, N. Andrews, and M. Edelstein, "Delivering Prenatal Pertussis Vaccine Through Maternity Services in England: What Is the Impact on Vaccine Coverage?" *Vaccine* 38, no. 33 (2020): 5332–5336, https://doi.org/10.1016/j.vaccine.2020.05.068.

8. E. Herrett, A. M. Gallagher, K. Bhaskaran, et al., "Data Resource Profile: Clinical Practice Research Datalink (CPRD)," *International Journal of Epidemiology* 44, no. 3 (2015): 827–836, https://doi.org/10.1093/ije/dyv098.

9. A. Wolf, D. Dedman, J. Campbell, et al., "Data Resource Profile: Clinical Practice Research Datalink (CPRD) Aurum," *International Journal of Epidemiology* 48, no. 6 (2019): 1740–1740G, https://doi.org/10.1093/ije/dyz034.

10. C. S. Osam, M. Pierce, H. Hope, D. M. Ashcroft, and K. M. Abel, "The Influence of Maternal Mental Illness on Vaccination Uptake in Children: A UK Population-Based Cohort Study," *European Journal of Epidemiology* 35, no. 9 (2020): 879–889, https://doi.org/10.1007/s10654-020-00632-5.

11. M. Peppa, S. L. Thomas, C. Minassian, et al., "Seasonal Influenza Vaccination During Pregnancy and the Risk of Major Congenital Malformations in Live-Born Infants: A 2010–2016 Historical Cohort Study," *Clinical Infectious Diseases* 73 (2020): e4296–e4304, https://doi.org/10.1093/cid/ciaa845.

12. A. Pottegård, L. C. Lund, Ø. Karlstad, et al., "Arterial Events, Venous Thromboembolism, Thrombocytopenia, and Bleeding After Vaccination With Oxford-AstraZeneca ChAdOx1-S in Denmark and Norway: Population Based Cohort Study," *BMJ* 373 (2021): n1114, https://doi.org/10.1136/bmj.n1114.

13. J. L. Walker, C. T. Rentsch, H. I. McDonald, et al., "Social Determinants of Pertussis and Influenza Vaccine Uptake in Pregnancy: A National Cohort Study in England Using Electronic Health Records," *BMJ Open* 11, no. 6 (2021): e046545, https://doi.org/10.1136/bmjopen-2020-046545.

14. S. L. Thomas, J. L. Walker, J. Fenty, et al., "Impact of the National Rotavirus Vaccination Programme on Acute Gastroenteritis in England and Associated Costs Averted," *Vaccine* 35, no. 4 (2017): 680–686, https://doi.org/10.1016/j.vaccine.2016.11.057.

15. N. Andrews, J. Stowe, G. Kuyumdzhieva, et al., "Impact of The Herpes Zoster Vaccination Programme on Hospitalised and General Practice Consulted Herpes Zoster in the 5 Years After Its Introduction in England: A Population-Based Study," *BMJ Open* 10, no. 7 (2020): e037458, https://doi.org/10.1136/bmjopen-2020-037458.

16. L. Smeeth, C. Cook, E. Fombonne, et al., "MMR Vaccination and Pervasive Developmental Disorders: A Case-Control Study," *Lancet* 364, no. 9438 (2004): 963–969.

17. J. L. Walker, N. J. Andrews, C. J. Atchison, et al., "Effectiveness of Oral Rotavirus Vaccination in England Against Rotavirus-Confirmed and All-Cause Acute Gastroenteritis," *Vaccine: X* 1 (2019): 100005, https://doi.org/10.1016/j.jvacx.2019.100005.

18. S. M. Shortreed, A. J. Cook, R. Y. Coley, J. F. Bobb, and J. C. Nelson, "Challenges and Opportunities for Using Big Health Care Data to Advance Medical Science and Public Health," *American Journal of Epidemiology* 188, no. 5 (2019): 851–861, https://doi.org/10.1093/aje/kwy292.

19. NHS Digital & UK Health Security Agency, "Childhood Vaccination Coverage Statistics—2020–2021," Published 2021, accessed October 12, 2021, https://digital.nhs.uk/data-and-information/publications/statistical/nhs-immunisation-statistics/england---2020-21.

20. Clinical Practice Research Datalink, "CPRD Aurum May 2022 Data Set. CPRD Aurum May 2022 (Version 2022.05.001) [Data Set]," https://doi.org/10.48329/t89s-kf12.

21. E. Herrett, S. L. Thomas, W. M. Schoonen, L. Smeeth, and A. J. Hall, "Validation and Validity of Diagnoses in The General Practice Research Database: A Systematic Review," *British Journal of Clinical Pharmacology* 69, no. 1 (2010): 4–14, https://doi.org/10.1111/j.1365-2125.2009.03537.x.

22. Medicines & Healthcare products Regulatory Agency, "Clinical Practice Research Data Link — Linked Data," Published 2021, https://www.cprd.com/linked-data.

23. UK Health Security Agency, "11 The UK Immunisation Schedule," in *Immunisation Against Infectious Disease (Green Book)*, ed. M. Ramsay (London: UK Health Security Agency, 2020).

24. T. Crocker-Buque, M. Edelstein, and S. Mounier-Jack, "A Process Evaluation of How the Routine Vaccination Programme Is Implemented at GP Practices in England," *Implementation Science* 13, no. 1 (2018): 132, https://doi.org/10.1186/s13012-018-0824-8.

25. Medicines & Healthcare Products Regulatory Agency, "Patient Group Directions: Who Can Use Them," Published 2017, accessed February 13, 2023, https://www.gov.uk/government/publications/patient-group-directions-pgds/patient-group-directions-who-can-use-them.

26. J. Matthewman, K. Andresen, A. Suffel, et al., "Checklist and Guidance on Creating Codelists for Electronic Health Records Research," *NIHR Open Research* 4 (2024): 20, https://doi.org/10.3310/nihropenres.13550.1.

27. A. M. Suffel, "GitHub Repository: Vaccine_methods_ehr," accessed November 20, 2023, https://github.com/Eyedeet/vaccine_methods_ehr_public.

28. UK Health Security Agency, *21 Measles: The Green Book*, ed. M. Ramsay (London: UK Health Security Agency, 2019).

29. P. Pesco, P. Bergero, G. Fabricius, and D. Hozbor, "Mathematical Modeling of Delayed Pertussis Vaccination in Infants," *Vaccine* 33, no. 41 (2015): 5475–5480, https://doi.org/10.1016/j.vaccine.2015.07.005.

30. J. Stowe, N. Andrews, S. Ladhani, and E. Miller, "The Risk of Intussusception Following Monovalent Rotavirus Vaccination in England: A Self-Controlled Case-Series Evaluation," *Vaccine* 34, no. 32 (2016): 3684–3689, https://doi.org/10.1016/j.vaccine.2016.04.050.

31. UK Health Security Agency, *24 Pertussis: The Green Book*, ed. M. Ramsay (London: UK Health Security Agency, 2016).

32. T. Waize, S. Anandarajah, N. Dhoul, and S. De Lusignan, "Variation in Clinical Coding Lists in UK General Practice: A Barrier to Consistent Data Entry?" *Journal of Innovation in Health Informatics* 15, no. 3 (2007): 143–150, https://doi.org/10.14236/jhi.v15i3.652.

33. UK Health Security Agency, *32 Tuberculosis: The Green Book*, ed. M. Ramsay (London: UK Health Security Agency, 2018), accessed November 14, 2022, https://www.gov.uk/government/publications/tuberculosis-the-green-book-chapter-32.

34. PRIMIS Team, "National Vaccinatio Uptake Programmes," accessed February 28, 2023, https://www.nottingham.ac.uk/primis/projects/national-vaccination-programmes.aspx.

35. Local Government Assocication, "Child Health Information Services," accessed November 14, 2022, https://www.local.gov.uk/topics/social-care-health-and-integration/public-health/children-public-health-transfer/child-health-information-services.

36. NHS England, "National Vaccination Programmes," accessed December 21, 2022, https://www.england.nhs.uk/contact-us/privacy-notice/national-flu-vaccination-programme/.

37. NHS Digital, "Quality and Outcomes Framework, 2021–2022," accessed November 16, 2022, https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data/2021-22.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.