Appendix 1 Methods for estimating patient level start and end dates in CPRD Aurum calculated by Boston Collaborative Drug Surveillance Program (BCDSP)

Table S1 provides the definitions of variables used to estimate a patient-level StartDate and EndDate in CPRD Aurum electronic medical records. *Table S2* describes the steps of the StartDate algorithm and *Table S3* describes the steps of the EndDate algorithm.

Table S1 Definitions

Term	Definition
Plausible range	Date range when electronic medical records would be possibly in use. Estimated StartDate and EndDate must fall in this range. Minimum = January 1, 1988 (GPs first started using computers for electronic medical records in 1988) or January 1 of Year of Birth, whichever came later Maximum = date of CPRD Aurum data download
regstartdate	CPRD Aurum data field for most recent registration with practice (See CPRD Aurum Data Specification). This field may be missing or may contain implausible values.
regenddate	CPRD Aurum data field for the date the patient's registration ended with the practice (See CPRD Aurum Data Specification). This field may be missing or may contain implausible values.
Rx	Prescriptions (Issue table) in Plausible range
Clin	Clinical information that indicates that a patient is active in the record, such as diagnoses, vaccinations, labs, diagnostic, clinical care, referrals in Plausible range. Use MedCodes (Observation Table) with EMISCatID 1-5, 7, 9, 11-12, 14-16, 20-21, 25, 27-29, 32-36, and 39-47 in CPRD Aurum MedCode dictionary.
FirstRecordDate	First date of Rx or Clin in Plausible range
LastRecordDate	Last date of Rx or Clin in Plausible range
FirstRxDate	First date of Rx in Plausible range
FirstClinDate	Last date of Clin in Plausible range
DeathDate	There are two death date fields in CPRD Aurum (See CPRD Aurum Data Specification). At the time of this study, emis_ddate was used as the death date for estimation of a patient EndDate. Future studies will use cprd_ddate rather than emis_ddate.
lcd	CPRD Aurum data field for last data collection date (practice-level) (See CPRD Aurum Data Specification)

Table S2 BCDSP's algorithm to estimate a patient-level StartDate in CPRD Aurum

Algorithm step	Primary condition	Secondary condition	Action
1	regstartdate missing		StartDate = FirstRxDate
2	regstartdate missing	If no Rx	StartDate = FirstClinDate
3	regstartdate missing	If no Rx or Clin	StartDate = missing
4	regstartdate < Jan 1, 2000		StartDate = FirstRxDate
5	regstartdate < Jan 1, 2000	If no Rx	StartDate = FirstClinDate on or after regstartdate
6	regstartdate < Jan 1, 2000	If no Rx or Clin	Set StartDate to missing*
7	regstartdate ≥ Jan 1, 2000	If difference between FirstRxDate and regstartdate is ≤365 days	StartDate = regstartdate
8	regstartdate ≥ Jan 1, 2000	If difference between FirstRxDate and regstartdate is >365 days	StartDate = last of regstartdate or FirstRxDate
9	regstartdate ≥ Jan 1, 2000	If no Rx	StartDate = FirstClinDate on or after regstartdate
10	regstartdate ≥ Jan 1, 2000	If no Rx or Clin	Set StartDate to missing*
11	If StartDate in plausible range cannot be calculated by this point		Set StartDate to missing*

Note: Patterns of recording by GPs in electronic records changed in the early 2000's due to technological advances and the introduction of Quality and Outcome Framework (QOF), a pay-for-performance scheme which improved the collection of data due to reporting requirements. *Patients for whom a plausible StartDate cannot be estimated do not have clinical or prescription records of high quality for use in research.

Table S3 BCDSP's algorithm to estimate a patient-level EndDate in CPRD Aurum

Algorithm step	Primary condition	Secondary condition	Action
1	StartDate is missing*		Set EndDate to missing*
2	DeathDate is not missing		EndDate = minimum, not zero, of DeathDate and lcd
3	regenddate is not missing and within plausible range		EndDate = minimum, not zero, of regenddate and lcd
4	regenddate is missing	If difference between lcd and LastRecordDate is ≤365 days	EndDate = lcd
5	regenddate is missing	If difference between lcd and LastRecordDate is >365 days	EndDate = earliest of lcd and LastRecordDate
6	EndDate is calculated	if EndDate < StartDate	EndDate = StartDate
7	EndDate in plausible range cannot be calculated		Set EndDate to missing*

*Patients for whom a plausible StartDate and/or EndDate cannot be estimated do not have clinical or prescription records of high quality for use in research.