CLAIMS

What is claimed is:

1. A method implemented at one or more computing machines comprising processing circuitry and memory, the method comprising:

accessing, at the processing circuitry, one or more medical data repositories storing data for a given patient from among a plurality of patients, the one or more medical data repositories storing pharmacy data, medical office visit data, and medical insurance transaction data;

identifying, based on one or more disease codes in the medical office visit data or the medical insurance transaction data, one or more biological conditions and a metastatic state for the given patient;

identifying, based on one or more drug codes in the pharmacy data, one or more lines of treatment for the given patient;

identifying, based on one or more insurance codes in the medical insurance transaction data, one or more medical procedures undergone by the given patient;

determining, based on a combination of the one or more biological conditions, the metastatic state, the one or more lines of treatment, and the one or more medical procedures, a primary diagnosis biological condition for the given patient;

assigning, based on the primary diagnosis biological condition, the given patient to a cohort of patients; and

providing an output representing the assigned cohort for the given patient.

2. The method of claim 1, wherein determining the primary diagnosis biological condition comprises:

creating, based on the medical insurance transaction data, a master gap table indicating intervals between two successive procedures for the given patient that have a same treatment name and insurance code, wherein the master gap table comprises columns for treatment name, insurance code, units, and gap length;

computing, based on the master gap table, a median gap table indicating the median gap for each combination of treatment name and insurance code, wherein the median gap table comprises columns for treatment name, insurance code, units, and gap length; and

determining the primary diagnosis biological condition based, at least in part, on data in the median gap table.

3. The method of any of claims 1-2, wherein the processing circuitry comprises a plurality of multithreaded graphics processing units (GPUs), the method further comprising:

determining, in parallel and using parallel threads of the plurality of multithreaded GPUs, assigned cohorts for multiple patients, including the given patient, from the plurality of patients.

4. The method of any of claims 1-3, wherein:

the disease codes comprise International Classification of Diseases (ICD) codes,

the drug codes comprise National Drug Code (NDC) codes, and

the insurance codes comprise Healthcare Common Procedure Coding System (HCPCS) codes.

5. The method of claim 4, wherein:

identifying, based on the one or more disease codes in the medical office visit data or the medical insurance transaction data, the one or more biological conditions for the given patient comprise: identifying that the given patient has lung cancer based on ICD codes associated with lung cancer.

6. The method of claim 5, wherein:

identifying, based on the one or more disease codes in the medical office visit data or the medical insurance transaction data, the metastatic state for the given patient is based on a secondary malignancy ICD code or HCPCS code.

7. The method of any of claims 1-6, wherein the primary diagnosis biological condition for the given patient is determined based on disease codes, drug codes or insurance codes associated with a date within a predefined date range.

8. The method of any of claims 1-7, wherein assigning the given patient to the cohort comprises:

analyzing one or more data tables that include medical insurance transaction data for the given patient, the one or more data tables being from the one or more medical data repositories;

determining one or more first insurance transactions that include one or more first code identifiers included in the one or more data tables, wherein the one or more first code identifiers correspond to diagnoses of the patient with respect to the one or more biological conditions;

determining one or more second insurance transactions that include one or more second code identifiers included in the one or more data tables, wherein the one or more second code identifiers corresponding to medical procedures administered with respect to the patient, the medical procedures being from the medical office visit data;

generating a medical headers table that includes a first number of columns storing the one or more first code identifiers, a second number of columns storing the one or more second code identifiers, and a plurality of rows with individual rows of the plurality of rows corresponding to a first medical insurance transaction of the one or more first insurance transactions or a second medical insurance transaction of the one or more second insurance transactions;

storing the medical headers table in the one or more medical data repositories; and

determining the cohort for the patient based on data in the medical headers table.

9. The method of claim 8, wherein:

individual first insurance transactions of the one or more first insurance transactions indicate a date of service of the individual first insurance transactions; and

individual second insurance transactions of the one or more second insurance transactions indicate a date of service of the individual second insurance transactions.

10. The method of claim 9, further comprising:

arranging the plurality of rows of the medical headers table in ascending order based on the dates of service such that a claim with an earliest date of service is a first row of the medical headers table and that a transaction with a most recent date of service is a last row of the medical headers table.

11. The method of any of claims 8-10, further comprising:

analyzing the one or more first code identifiers to determine that a first code identifier of the one or more first code identifiers is included in a group of insurance code identifiers that correspond to the one or more biological conditions.

12. The method of claim 11, wherein:

the first code identifier is arranged according to a first format that corresponds to a first classification of insurance code identifiers;

the first classification of insurance code identifiers corresponds to International Classification of Diseases version 9 (ICD-9).

13. The method of any of claims 11-12, wherein:

the first code identifier is arranged according to a second format that corresponds to a second classification of insurance code identifiers;

the second classification of insurance code identifiers corresponds to International Classification of Diseases version 10 (ICD-10).

14. The method of any of claims 11-13, wherein:

the one or more biological conditions include a plurality of subtypes;

individual subtypes of the plurality of subtypes correspond to a subset of the group of insurance code identifiers that correspond to the one or more biological conditions; and

the method further comprises determining that the first code identifier is included in a first subset of the group of insurance codes that corresponds to a first subtype of the biological condition.

15. The method of any of claims 11-14, wherein the biological condition is cancer and the plurality of subtypes include at least one of lung cancer, breast cancer, or colorectal cancer.

16. The method of claim 15, further comprising:

determining one or more third insurance transactions having a date of service that is within a predefined time period ending on a first date of service;

analyzing one or more third insurance code identifiers of the third insurance transactions with respect to the first code identifiers and the second code identifiers.

17. The method of claim 16, further comprising:

determining that the one or more third insurance code identifiers are not included in the first code identifiers and the second code identifiers; and

determining, based on the third insurance code identifiers, that the patient is included in a cohort of patients in which a given subtype of the one or more biological conditions is present.

18. The method of any of claims 16-17, wherein the one or more third insurance code identifiers correspond to an additional biological condition.

19. The method of any of claims 16-18, further comprising:

determining that the one or more third insurance code identifiers are included in a portion of the group of insurance code identifiers that are not included in the subset of the group of insurance code identifiers;

determining that a date of service of at least one of the one or more third insurance transactions is a same date as the date of service as one of the one or more first insurance transactions;

determining that there are no other additional insurance transactions having insurance code identifiers included in the group of code identifiers; and

determining that the patient is included in a cohort of patients in which the subtype of the biological condition is present.

20. The method of any of claims 16-19, further comprising:

determining that the one or more third insurance code identifiers are included in a portion of the group of insurance code identifiers that are not included in the subset of the group of insurance code identifiers;

determining that a date service of at least one of the one or more third insurance claim is prior to the date of service of one of the one or more first insurance transactions and within the predefined time period; and

determining that the patient is not included in a cohort of patients in which the subtype of the biological condition is present.

21. A method implemented at one or more computing machines comprising processing circuitry and memory, the method comprising:

accessing, at the processing circuitry, one or more medical data tables storing medical insurance transaction data for a plurality of patients, the one or more medical data tables comprising a date column and a diagnosis column;

identifying, using the processing circuitry and based on the diagnosis column, a set of patients having a specified biological condition, the set of patients being from among the plurality of patients;

determining, for each patient in the set of patients, an earliest date when the patient received a diagnosis of the specified biological condition;

identifying, using the processing circuitry and based on the diagnosis column and the date column, a cohort of patients from among the set of patients, the cohort of patients lacking a diagnosis from a collection of biological conditions associated with a date occurring during a predefined time window before the earliest date when the patient received the diagnosis of the specified biological condition; and

providing an output representing the cohort.

22. The method of claim 21, wherein the diagnosis column stores International Classification of Diseases version 9 (ICD-9) or International Classification of Diseases version 10 (ICD-10) codes.

23. The method of any of claims 21-22, wherein the specified biological condition is lung cancer, wherein the collection of biological conditions comprises cancers different from lung cancer, wherein the predefined time window before the earliest date is six months before the earliest date.

24. The method of any of claims 21-23, wherein the specified biological condition is a specified type of cancer, the method further comprising:

determining a metastatic state of at least one patient from the cohort.

25. The method of claim 24, wherein the metastatic state is determined based on a secondary malignancy International Classification of Diseases (ICD) code or Healthcare Common Procedure Coding System (HCPCS) code.

26. The method of any of claims 21-25, wherein identifying the cohort comprises:

arranging rows associated with patients in the set by date; and

accessing rows associated with the predefined time window to identify patients in the set that lack the diagnosis from the collection of biological conditions during the predefined time window.

27. A system comprising:

processing circuitry; and

a memory storing instructions which, when executed by the processing circuitry, cause the processing circuitry to perform operations comprising:

accessing, at the processing circuitry, one or more medical data repositories storing data for a given patient from among a plurality of patients, the one or more medical data repositories storing pharmacy data, medical office visit data, and medical insurance transaction data;

identifying, based on one or more disease codes in the medical office visit data or the medical insurance transaction data, one or more biological conditions and a metastatic state for the given patient;

identifying, based on one or more drug codes in the pharmacy data, one or more lines of treatment for the given patient;

identifying, based on one or more insurance codes in the medical insurance transaction data, one or more medical procedures undergone by the given patient;

determining, based on a combination of the one or more biological conditions, the metastatic state, the one or more lines of treatment, and the one or more medical procedures, a primary diagnosis biological condition for the given patient;

assigning, based on the primary diagnosis biological condition, the given patient to a cohort of patients; and

providing an output representing the assigned cohort for the given patient.