CLAIMS

What is claimed is:

1. A method comprising:

generating, by a computing system including processing circuitry and memory, a data file including first tokens generated using a first hash function, individual first tokens corresponding to a respective individual of a group of individuals having data stored by a molecular data repository;

sending, by the computing system, the data file to a health insurance claims data management system;

obtaining, by the computing system and from the health insurance claims data management system, in response to the data file, health data corresponding to the group of individuals;

generating, by the computing system, a number of identifiers using a second hash function that is different from the first hash function, each identifier corresponds to one or more tokens related to each individual of the group of individuals;

obtaining, by the computing system and using the number of identifiers, second data from the molecular data repository for the group of individuals;

determining, by the computing system, respective portions of the first data that correspond to respective portions of the second data for the group of individuals; and

generating, by the computing system, an integrated data repository that stores the respective portions of the first data and the respective portions of the second data in relation to respective identifiers of the number of identifiers.

2. The method of claim 1, comprising:

determining, by the computing system, a first set of data processing instructions that are executable in relation to first data stored by the integrated data repository;

causing, by the computing system, the first set of data processing instructions to be executed to analyze first health insurance claims codes included in the first data to determine a first subset of the group of individuals in which a biological condition is present; and

generating, by the computing system, a first dataset indicating the subset of the group of individuals in which the biological condition is present.

3. The method of claim 2, comprising:

determining, by the computing system, a second set of data processing instructions that are executable in relation to second data stored by the integrated data repository;

causing, by the computing system, the second set of data processing instructions to be executed to analyze the second health insurance claims codes included in the second data to determine one or more treatments provided to a second subset of the group of individuals; and

generating, by the computing system, a second dataset indicating the one or more treatments provided to the second subset of the group of individuals.

4. The method of claim 3, comprising:

determining, by the computing system, a third subset of the group of individuals that includes a portion of the first subset of the group of individuals that overlaps with a portion of the second subset of the group of individuals;

receiving, by the computing system, a request to perform an analysis of the first dataset and the second dataset in relation to the third subset of the group of individuals; and

analyzing, by the computing system and in response to the request, the first dataset and the second dataset with respect to the third subset of the group of individuals to determine a measure of significance of a characteristic of the third subset of the group of individuals with respect to the biological condition.

5. The method of claim 4, comprising:

determining, by the computing system, one or more genomic mutations present in the third subset of the group of individuals;

determining, by the computing system, a plurality of treatments provided to the third subset of the group of individuals; and

determining, by the computing system, respective survival rates for the third subset of the group of individuals.

6. The method of claim 5, wherein the measure of significance corresponds to survival rate

with respect to a treatment of the plurality of treatments and a genomic mutation of the one or more genomic mutations.

7. The method of claim 6, comprising determining, by the computing system and based on measure of significance, an effectiveness of the treatment for the third subset of the group of individuals.

8. The method of claim 7, comprising determining, by the computing system, individuals in third subset of the group of individuals that have not received the treatment.

9. The method of claim 8, comprising administering one or more therapeutically effective amounts of the treatment to the individuals in the third subset that have not received the treatment.

10. The method of any one of claims 1-9, wherein:

the integrated data repository is arranged according to a data repository schema that includes a plurality of data tables and a plurality of logical links between the plurality of data tables;

individual logical links of the plurality of logical links indicating one or more rows of a data table of the plurality of data tables that correspond to one or more additional rows of an additional data table of the plurality of data tables.

11. The method of claim 10, wherein the plurality of data tables include:

a first data table that stores genomics data of the group of individuals;

a second data that stores data related to one or more patient visits by individuals to one or more healthcare providers;

a third data table that stores information corresponding to respective services provided to individuals with respect to one or more patient visits to one or more healthcare providers indicated by the second data table;

a fourth data table that stores personal information of the group of individuals;

a fifth data table that stores information related to a health insurance company or

governmental entity that made payment for services provided to the group of individuals;

a sixth data table storing information corresponding to health insurance coverage information for the group of individuals; and

a seventh data table that stores information related to pharmaceutical treatments obtained by the group of individuals.

12. The method of any one of claims 1-11, wherein the number of identifiers generated using the second hash function comprise intermediate identifiers; and the method comprises:

applying, by the computing system, a salt function to the intermediate identifiers to generate a final set of identifiers.

13. The method of any one of claims 1-12, comprising:

obtaining, by the computing system, information from an additional data repository that includes electronic medical records of an additional group of individuals;

determining, by the computing system, a subset of the additional group of individuals that corresponds to the group of individuals having data stored by the genomics data repository; and

modifying, by the computing system, the integrated data repository to store at least a portion of the information of the medical records of the subset of the additional group of individuals in relation to the number of identifiers.

14. The method of claim 13, comprising:

performing, by the computing system, one or more optical character recognition operations with respect to the additional information;

analyzing, by the computing system, the additional information obtained from the additional data repository to determine one or more portions of the additional information to remove to produce a corpus of information.

15. The method of claim 14, comprising:

analyzing, by the computing system, the corpus of information to determine a portion of the subset of the additional group of individuals that correspond to one or more biomarkers; and

generating, by the computing system, one or more data structures that store identifiers of

the portion of the subset of the additional group of individuals and that store an indication that the portion of the subset of the additional group of individuals corresponds to the one or more biomarkers.

16. The method of claim 15, comprising:

storing, by the computing system, the one or more data structures in an intermediate data repository;

performing, by the computing system, one or more de-identification operations with respect to the identifiers of the portion of the subset of the additional group of individuals before modifying the integrated data repository to store at least a portion of the additional information of the medical records of the portion of the subset of the additional group of individuals in relation to the number of identifiers.

17. The method of any one of claims 1-16, wherein the molecular data repository stores at least one or more of genomic information, genetic information, metabolomic information, transcriptomic information, fragmentomic information, immune receptor information, methylation information, epigenomic information, or proteomic information.

18. A system comprising:

one or more hardware processing units;

one or more computer-readable storage media storing computer-executable instructions that, when executed by the one or more hardware processing units, cause the system to perform operations comprising:

generating a data file including first tokens generated using a first hash function, individual first tokens corresponding to a respective individual of a group of individuals having data stored by a molecular data repository;

sending the data file to a health insurance claims data management system;

obtaining, from the health insurance claims data management system, in response to the data file, health insurance claims data corresponding to the group of individuals;

generating a number of identifiers using a second hash function that is different

from the first hash function, each identifier corresponds to one or more tokens related to each individual of the group of individuals;

obtaining, using the number of identifiers, second data from the molecular data repository for the group of individuals;

determining respective portions of the first data that correspond to respective portions of the second data for the group of individuals; and

generating an integrated data repository that stores the respective portions of the first data and the respective portions of the second data in relation to respective identifiers of the number of identifiers.

19. One or more non-transitory computer-readable storage media storing computer-executable instructions that, when executed by one or more hardware processing units, cause the system to perform operations comprising:

generating a data file including first tokens generated using a first hash function, individual first tokens corresponding to a respective individual of a group of individuals having data stored by a molecular data repository;

sending the data file to a health insurance claims data management system;

obtaining, from the health insurance claims data management system, in response to the data file, health insurance claims data corresponding to the group of individuals;

generating a number of identifiers using a second hash function that is different from the first hash function, each identifier corresponds to one or more tokens related to each individual of the group of individuals;

obtaining, using the number of identifiers, second data from the molecular data repository for the group of individuals;

determining respective portions of the first data that correspond to respective portions of the second data for the group of individuals; and

generating an integrated data repository that stores the respective portions of the first data and the respective portions of the second data in relation to respective identifiers of the number of identifiers.