Appendix A: Validation

I. A Co-occurrence Probability Measure for comparing Model Populations

Here we describe our approach to comparing two populations of patients with respect to the statistical distributions of a set of discrete variables. The comparisons will be between real versus simulated populations, and between simulated populations generated before and after adding our module to the simulator. Discrete variables include categorical demographic attributes like patient age or gender, as well as SNOMED condition codes. We call the method described here "co-occurrence probability measure" (COPM) because it is based on the frequency with which these variables appear together in the same patient encounter. An explainable, quantitative population-level evaluation function such as this is instrumental for measuring progress in modelling not only to guide the present study, but as a contribution to the future development of Synthea.

An evaluation measure should meet these criteria:

- It measures divergence—a positive number that is zero just for identical populations.
- The measure is additive. The aggregate metric can be explained as the contribution of individual comparisons, so it remains meaningful when applied to interesting subsets of variables.
- It is a well-behaved mathematical function. The function should be smooth, so that small difference in probabilities lead to small differences in the function value and convex , so that it has one minimum value, equal to zero, that is reached when the populations are statistically identical.
- It is applicable to the co-occurrence statistics and derivable from probabilistic Bayes network models used in this study.

We evaluate a population, either from a simulation or directly from data, by computing pairwise co-occurrence counts of variable values over encounters and use these to compute a matrix of pair-wise conditional probabilities between variables (for example, of all the people who have hypertension, how many also have coronary artery disease?). For *N* variables these form an *N*-by-*N* "Conditional Probability Matrix (CPM) for each population. Population matrices are then compared by Kullback–Leibler divergence, an entropy-based measure. This divergence measure can be calculated exactly, and, after making some reasonable assumptions, will be shown to have the desired mathematical properties.

II. Representing a Simulation as a Probability Model

The output of a simulation by Synthea is a sequence of medical *encounters* over time for each patient in a population. Each encounter contains a list of disease codes describing the patient state at that instant. Information about the encounter is captured in a set of discrete variables. For medical condition codes (like hypertension) these are binary, but in some cases the variable is multi-valued such as for patient's age category, or to represent a state of disease progression (e.g. the stages of chronic kidney disease, CKD, are encoded as integers 0 through 5). As described in our report, substantial effort has been invested in creating a consistent mapping between coding taxonomies used by Synthea (SNOMED concepts) and the medical record database used in this study (coded in ICD10) so that corresponding concepts can be compared.

This allows us to reformat the simulation data into a table with a column for each discrete variable, and a row for each patient encounter. From this we can estimate the joint probability distribution between variables. The full joint probability table of cross tabulations among N variables consists of over 2^N entries; since our data contains hundreds of variables, this would be an unmanageably complex, though extremely sparse matrix. Bayes networks provide a well-known approach to manage this complexity by approximating the full joint probability table in a compact, computationally manageable network representation. Given the locality assumptions of the network---that each variable is conditioned only on a few, say no more than k other variables, the total number of parameters in the model reduces to order $O(2^{k+1} N)$.

III. Reducing a Probability Model to a Comparable Representation

There are two ways to obtain the pair-wise conditional probabilities, either by reducing a Bayes network learned from the data, or from the co-occurrence counts generated for item-set analysis. Both methods give essentially equivalent results but are computed differently and offer some different options.

Because the Bayes network reduces the joint probability over the full set of variables to a concise model of the data, it can be put to various uses. The structure of the network supports some causal claims among variables, which we apply to support findings about disease co-morbidity, as illustrated in the report's Bayes network for chronic kidney disease. Exact, fixed time algorithms are available to compute the conditional probability of any set of variables M_i conditioned on any exclusive set of variables C_j, to obtain $P(M_i \mid C_j)$. Computing pairwise conditionals over variables of interest by reducing the Bayes network is an one way to compute the conditional probability matrix used for validation.

For item-set computations, encounter disease codes are tabulated to get pair-wise counts of co-occurrences directly from the encounter data whenever two or more codes occur in the same encounter. Using the total counts of each item in the pair, we calculate conventional "item-basket" [Hastie et al. 2001] statistics, notably "confidence" that corresponds to the "plug-in" conditional probability estimate of one item conditioned on another. This generates an $N(N-1)$ matrix of distinct entries, since conditioning an item on itself has probability 1 and can be ignored.

The validation results presented in the report use item-set computations to generate the CPM. The advantages of learning Bayes networks from the data and then computing the CPM are discussed subsequently.

IV. Choice of a Probabilistic Measure of Similarity between Conditional Probabilities

We considered two approaches, one involving differencing proportions using a variation of the Chi-squared test, the other derived from Kullback–Leibler divergence (KL). [Cover, 1991].

Chi-squared tests whether multinomial data deviate significantly in distribution, and it can be generalized to compare observed proportions across groups. Presumably we can compare cells between the two matrices pair-wise, resulting in a Chi-squared statistics with *N(N-1)* terms. Aside from the violation of the assumption that each cell is conditionally independent given the model parameters, it leads to the question of how to estimate the model's degrees of freedom---a large number that is certainly of the order of the number of cells. Even if this was not an issue, the test on such a large set almost certainly generates such a large statistic as to give infinitesimal P-values. This is not surprising since in the limit of large data the test almost certainly will register a significant difference between the two distributions.

Using a similar basis for comparison we can apply KL divergence, or relative entropy to the pair of probabilities in corresponding cells in the two matrices. Averaging this value over all cells obtains:

$$\mathrm{KL}(\mathrm{P} \parallel Q) = \frac{1}{N(N-1)} \sum\nolimits_{i,j,i\neq j} [\mathsf{P}(P_{i,j}) \log \frac{\mathsf{P}(P_{i,j})}{\mathsf{P}(Q_{i,j})} + (1-\mathsf{P}(P_{i,j})) \log \frac{1-\mathsf{P}(P_{i,j})}{1-\mathsf{P}(Q_{i,j})}]$$

Where *P* is the "reference" matrix of conditional probabilities, and *Q* is the "test" matrix. The two terms inside the summation are the KL divergence for one matrix element, considered as a binary distribution. These entropy-like terms are be summed, recognizing that this is an approximation due to the lack of independence among cells, to obtain an average cell value. KL divergence is proportional to the number of "bits" to represent the difference when using natural logarithms.

Expanding KL divergence in a Taylor series reveals that the first couple terms of Chi-squared statistic approximates the KL divergence. Use of KL divergence is sometimes re-labelled the "G-test" [McDonald, J.H. (2014)].

It can be shown that the properties of KL divergence meet the criteria proposed in the introduction.

V. Evaluation of A Synthea Simulated Population with an actual Population

As covered in the report, we showed how it is possible to improve Synthea's co-morbidity modelling by deriving probability models from clinical records. The degree of improvement has been validated by comparing the simulation data generated by the modified Synthea model to data from clinical records using COPM.

The validation followed these steps.

The schemas of both simulated data and the clinical encounter data are the same. For the clinical data much attention was paid to developing a mapping from the clinical record ICD10 codes to the corresponding SNOMED codes used by Synthea. This artifact of the study will be valuable for future improvements to Synthea, but it is not part of the model validation. The mapping chose a subset of SNOMED codes that regularly appear in Synthea modules, numbering over 200.

Initial data is in the form of encounters; a record of a patient's clinical visit, either in hospital or for ambulatory care. Each patient encounter includes a list of medical codes. These are processed analogously to a "market-basket" of items to generate a table of all pairs of code occurrences. If the encounter record contains only one code, then it generates no co-occurrence record, but in the typical case, one encounter record with numerous codes generates all pair-wise combinations of co-occurrences. This is an approximation that overlooks higher-order interactions, but we feel is adequate to find the relations among diseases we are looking for.

The next step counts all item-pair combinations over all encounters and patients to create two tables of length of the number of SNOMED codes squared, for each data source. The tables contain conventional item-set statistics, of which the "confidence" corresponds to an empirical estimate of conditional probability of item1 given item2. These tables are the source of the co-occurrence visualizations in the study that are used to explore overall properties of the data.

Even with the populations of 100,000 used in this study, there will be zero counts for some item-pair combinations. These should not be ignored, since they are neither missing values, nor impossible states. Leaving as zero is also inappropriate as an estimate of their probability and would lead to division by zero in probability divergence measures. In this case we "smooth" the counts by adding 5 to both numerator and denominator of the estimates of conditional probability, corresponding roughly to a prior probability of roughly one occurrence per the number of patients in the dataset.

The "long" tables of item-pair conditional probabilities need to be pivoted to "wide" form, with item1 along the rows and item2 along columns, so that the row and column labels are the same. For comparison purposes we remove variables for demographic and risk factors to focus the comparison on the effects of one disease state on another. We consider differences in demographics to be second order for the effects under study. This is a reasonable approximation for this study, but the proper way to compensate for demographic differences between samples by using a Bayes network model to estimate the conditional probabilities is discussed in the conclusions, and in principle could be part of COPM.

In the last preparation step, the two CPM matrices are trimmed to just include common variables. This underestimates the difference between samples, since the missing variables would be assigned a small constant value, resulting in a large contribution to the COPM measure. The resulting matrices included 103 variables. The pair of CPM matrices is the input to the KL divergence measure.

The data reduction of Synthea and clinical data were performed on different platforms to avoid any possibility of leaking confidential data. The final CPM matrices to not pose a confidentiality risk.

VI. Sources of Approximations, and ways to improve COPM.

Demographic bias

In principle, as mentioned, COPM measures the divergence between disease state distributions at a point in time. A Bayes network learned from the data directly approximates this distribution, capturing not only pair-wise but higher order interactions in the data. In concept one can then query the Bayes network for the required conditional probabilities. Since different samples may have different demographic distributions, the Bayes network has the advantage that the demographic distributions learned from one data sample may be replaced by a common distribution for both samples, which would remove the effect of population bias on the CPM. This improvement due to the causal nature of Bayes networks. Such a correction will be valuable in future applications of COPM, but by use only of the co-occurrence statistics we did not take advantage of it in the present study.

Acceptance selection bias

There are biases to be accounted for in the actual data. The well-known Berkson's paradox [Pearl (2018) p 197.] is present as a consequence simply of the nature of encounter patient visits. In short, one observes for any pair of disease, both the occurrence of either, or the combination of the pair, but not the lack of occurrence of both, since the last case does not result in an encounter. The result in probability is an apparent negative correlation among the pair of states when no true correlation exists. Due to the nature of the simulation the same property is true of both simulated and population data, so the accuracy of the COPM measure is not called into question. However, to estimate true co-morbidity rates in the population the missing observations would need to be accounted for.

Bias due to age

We do use patient age in the Bayes network model. It affects the incidence rate of diseases broadly, as shown by the range of disease conditioned by patient age, either directly or indirectly in the Bayes network. Since both the simulated and actual populations contain a distribution of patient ages, we can make the reasonable assumption that the population age distribution does not vary significantly over the time interval studied and both populations are assumed to fit stationary probability models for disease co-occurrence. Age bias, similar to demographic biases can be adjusted for by changing the conditioning in the Bayes network.

Temporal bias

Notably we are applying a measure at an instant in time ("synchronic") when conceivably a measure that takes into account evolution over time ("diachronic") would be called for, such as proposed by other authors such as [Marini, 2015]. A diachronic approach requires estimation of transition probabilities – also known as "hazard rates" [Clarke, 2013]. The matrix of such transition probabilities can be assumed to form a discrete time Markov chain. Given a model of transition probabilities, assuming some regularity conditions on the Markov chain, and a starting distribution, one can compute the time evolution of the fraction of the population in each disease state. Such Markov chains possess a final equilibrium distribution. What we observe in our CPM at a point in time is not this equilibrium, but a stationary state that is a combination of the population age distribution, its effect on disease onset, and the transition probability-based disease progression. We make the strong, but reasonable assumption that the synchronic CPM we observe represents a stationary distribution as a combination of these factors. The obvious condition where this is not the case is during a pandemic, as has been recently experienced, where disease incidence and death rates cause secular change in the observed distributions.

VII. Why COPM should be adopted by Synthea.

Science advances when there is an agreed-upon criterion to distinguish a better result from others. Commonly statistical methods are relied on when results are uncertain. The COPM measure of similarity is a general tool for measuring the accuracy of simulations against actual data. We've shown the COPM measure can be applied when comparing events from population simulation models with actual data.

The method makes no assumption about the source of the statistics used for comparison: It can compare two population data sets either at an instant in time, or over a time span, assuming the generating process is stationary. Alternately it can compare statistics from data with probability models with such as a Bayes network, making comparisons possible without the need to generate a simulation data set.

In this study we demonstrated the process to make an incremental improvement to a Synthea module derived from a Bayes network model, then used COPM to validate the degree of improvement when compared to actual population data.

In a related area there are current ethical fairness concerns about quantitative models when applied to health policy choices. In population simulation work biases in demographics due to gender, race, or age imbalances in data used to build the model can result in biased choices. Notably the Bayes network can correct for such bias by adjusting the model to generate results from one demographic distribution while having been built with data having a different distribution. As described, one removes the current demographic distribution and replaces the distribution of these nodes in the Bayes network to meet the new circumstances. The model's modified distribution of diseases can then be validated against actual data using COPM.

This is a general advantage of working with a model that comprises the full joint probability. distribution.

References

Clarke C.L. et al. "Natural History of Left Ventricular Ejection Fraction in Patients with Heart Failure" (2013) *Circ Cardiovasc Qual Outcomes* DOI: 10.1161/CIRCOUTCOMES.111.000045.

Cover, T. M. and J. A. Thomas, (1991) *Elements of Information Theory,* (New York, Wiley).

Hastie, T., R. Tibshirani, J. Friedman, (2001), *The Elements of Statistical Learning,* (New York, Springer).

Marini, S, Trifoglio E, Barbarini N, Sambo F, Di Camillo B, Malovini A, Manfrini M, Cobelli C, Bellazzi R., (2015), "A Dynamic Bayesian Network model for long-term simulation of clinical complications in type 1 diabetes," *Journal of Biomedical Informatics,* Volume 57, pp 369-376.

McDonald, J.H. (2014). "G–test of goodness-of-fit". *Handbook of Biological Statistics* (Third ed.). Baltimore, Maryland: Sparky House Publishing. pp. 53–58.

Pearl J and MacKenzie D. (2018) *The Book of Why - the New Science of Cause and Effect.* (New York , Basic Books).