We have now achieved MCMC convergence in our LTC model with all parameters achieving an R-hat (also known as the Gelman-Rubin convergence diagnostic, Gelman and Rubin 1992, Brooks and Gelman 1997) below 1.2 (see figure below). The results of the original and converged model differed very little due to the steps taken in the original analysis, with tiny differences in the final estimates of Years of Life Lost. The multimodality and extreme flexibility of our original model led to poor converge convergence. By restricting the model a small amount through more slightly informative priors, changing how the correlation matrix was samples, and treating the absence of a diagnosis in the individual patient data as absence of the disease at a clinically significant level we were able to dramatically improve MCMC mixing and thus convergence. More details on these steps and why convergence did not substantially change the estimate of Years of Life Lost are explained below. We favoured achieving convergence for the more complex model as simpler approaches, such as the sensitivity analysis or the incorporation of data from other diseases necessarily reduce biological realism (and hence introduce bias). Unfortunately, the data were not available to meaningfully incorporate age distribution into the LTC joint prevalence model as the data are aggregated separately. Technically, this is achievable by incorporating the age structure as priors but with the lack of additional information in the model the prior would be returned giving an equivalent model (but perhaps with worse convergence) to the two stage approach we did.

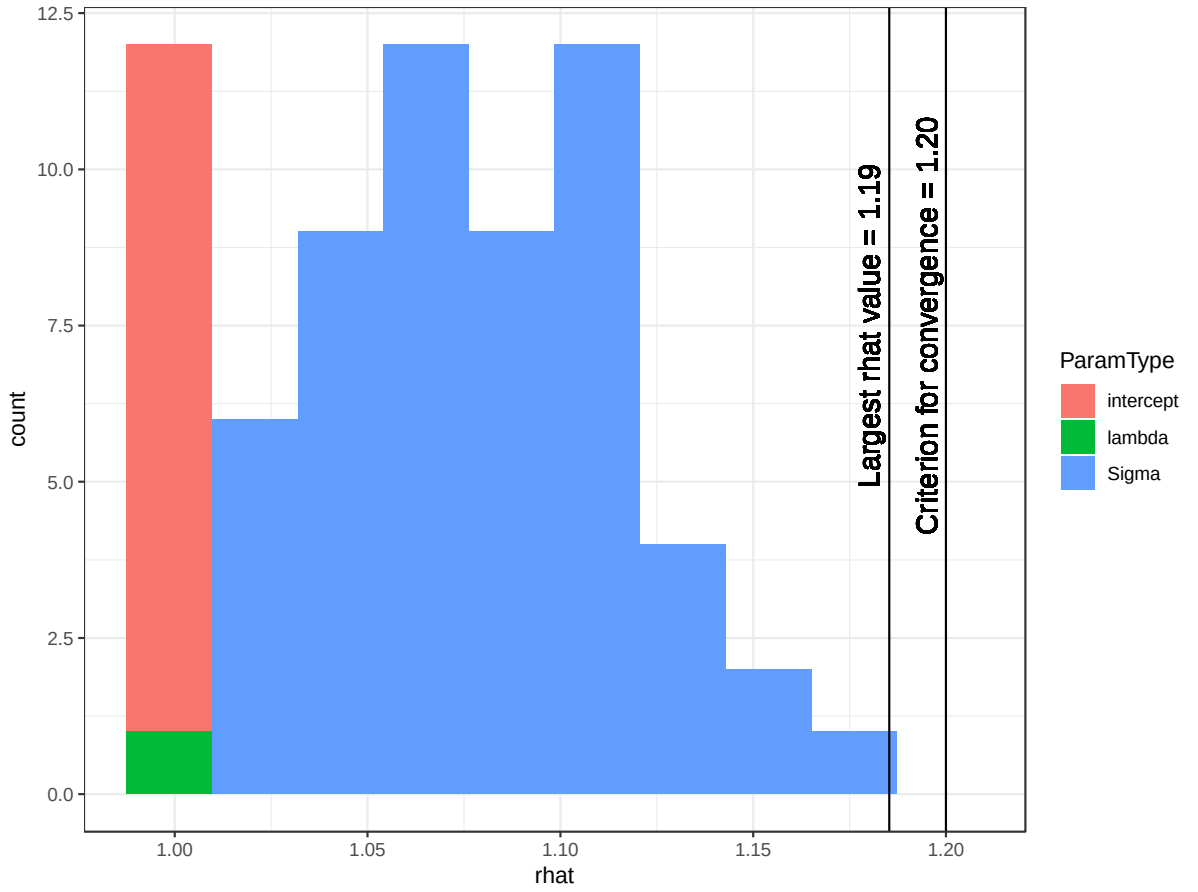

Figure 1 Distribution of R-hat (Gelman-Rubin statistic) values for each parameter. Colour coded by parameter type. Intercept parameters estimate the frequency of each disease in the population (corresponding to the column sums in our model), lambda is the rate parameter of the Poisson distribution used to impute comorbidity counts (corresponding to the

row sums in our model) and the Sigma parameter is the correlation parameter between the diseases (this is estimated in the MCMC as a precision and then rescaled post-MCMC sampling, the rhat values are for the rescaled samples).

Details on model modifications to improve convergence

The first issue we identified in the MCMC mixing of our original model was the sampling of the correlation matrix. We viewed a correlation matrix formulation as essential as the variance and mean parameters are not identifiable in the multivariate probit model (see Equation 1 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2966284/), thus by using a correlation matrix (with fixed variances of 1) the variance is fixed, making the mean identifiable. As there are no correlation matrix priors available in JAGS, to sample correlation matrices we had to choose between a computationally very expensive formulation where we sampled from an unconstrained precision matrix from a Wishart prior, inverted it and renormalised it each iteration, or using a non-conjugate prior that we could constrain to produce correlation matrices but that sampled the components of the matrix separately. As matrix inversions and normalisations are computationally very expensive, we opted for the latter approach, sampling the correlation matrix components from independent beta distributions. This imposes two limitations. Firstly, the MCMC chains are not restricted to sampling positive definite matrices leading to many samples being rejected. Secondly, the correlation matrix space is constrained leading to MCMC chains being limited in their ability to move between local optima.

These limitations prevented our original model mixing well, so we developed an alternative sampling strategy. In the new version of the model, we use a Wishart prior to sample precision matrices directly, but, crucially, we apply the expensive corrections to make the model identifiable after sampling. This allows us to sample quickly and maintain the flexibility of moving freely through covariance space, allowing the samplers to move between local optima. However, this strategy also opens the parameter space infinitely so some constraint needs to be added. We therefore used a scaled Wishart matrix with a scale of 1 and 50 degrees of freedom. This places a half-t prior with the same parameters on the variance, a slight shrinkage prior on the correlations (NEED TO PUT GRAPH SOMEWHERE?), equivalent to drawing 50 additional samples from an identity matrix for inclusion in our analysis. Similarly, we placed a slightly more restrictive prior on the mean values (a normal with mean zero and precision of 0.2) which is uninformative across the domain of a probit and still only weakly within the tolerance of the half-t prior on the variance described above.

We also slightly relaxed the precision by which the imputed row and column sums are linked to the comorbidity counts and disease frequencies, respectively. Originally, we specified that the difference between the imputed patient values and the aggregate Italian data would be normally distributed with a mean of 0 and standard deviation of 0.001 (precision of 100). This approach allows almost no deviation between the two which again strongly constrains the MCMC sampling, and is also biologically unrealistic as it implicitly assumes no patient had a missed or misdiagnosed disease. We relaxed these constraints by allowing a standard deviation of 0.25 for the row sums and comorbidity count, and 5 for the column sum and disease count equalities. Across the data, these allow for small differences between the observed values and imputed values, and much improved MCMC mixing.

Finally, we slightly modified our use of the two data types. In our original model we treated the individual patient data as "presence-only" disease data, rather than treating the absence of a disease diagnosis as an absence of the disease. This made these data less informative and also introduced a discrepancy in how the aggregated and individual data were being treated. In our updated model, absence of a diagnosis is treated as absence of the disease (i.e. as a 0). The added flexibility in the row and column sums above allows for errors in these assignments to be accommodated. The original model also included patients with no long term conditions, however, we excluded these patients from the second analysis to focus exclusively on patients with long term conditions.

Together, these modifications have allowed us to achieve model convergence without sacrificing biological reality. Fortunately, as our original approach appropriately represented the uncertainty in our approximations, the estimated YLLs have not changed as a result of these modifications. As outlined above, realistic simplifications of the model are not possible and due to the lack of covariate data model selection was not possible.

Why the estimates do not differ

Lack of convergence in our original model, which persisted despite our extensive numerical runs of the MCMC, was generated by multimodalities in the posterior distribution that were being explored by the MCMC chains (i.e. the MCMC could find several equally plausible explanations for the observations). Fortunately, our MCMC chains covered the parameter space well and thus provided an approximation of this (multimodal) target distribution. We represented this uncertainty by taking large samples from all the MCMC chains. This propagated all uncertainty on to our final estimates of YLL, meaning that our results were inclusive of any ambiguities due to MCMC identifiability.

Note that statistical inference in complex models makes approximations, like adopted by us, inevitable due to its intrinsic computational costs. Even if only partially converged, our method fared no worse than the converged model or than the established alternative approximation procedures from the computational inference literature. An insufficiently converged MCMC chain may only represent a particular mode. This is conceptually similar to modal approximations, like the Laplace approximations, which are widely accepted in the Statistics literature (e.g. Gelman et al., Bayesian Data Analysis, 3rd edition, chapter 13). However, the Laplace approximation makes a particular distributional assumption (multivariate Gaussian), which our insufficiently converged MCMC chains avoid, thus reducing any intrinsic estimation bias (see Ferguson et al., Applied Statististics (2017) 66, Part 4, pp. 869–890, Section 7 for more details on this). Moreover, our combination of local MCMC simulations associated with different modes was conceptually similar to the method of "Approximate Bayesian ensembling (ABE)"; see e.g. Pearce, Leibfried, Brintrup, Zaki and Neely, Proc. AISTATS 2020. ABE is an ensemble of parameters obtained from repeated local optimisations; our method is an ensemble of local MCMC chains. While conceptually similar, the exploration of the parameter space around the local optima afforded by our method reduced the risk of uncertainty underestimation that may affect ABE. In summary, despite its approximate nature, our ensemble of local (and hence incompletely converged) MCMC simulations was of the same standard as alternative approximate inference methods from

the computational inference literature that have to be adopted when running MCMC simulations to convergence becomes computationally infeasible.