How much multi-morbidity is due to old age and established risk factors?

Anthony J. Webster

Nuffield Department of Population Health, Lindgren Group, Big Data Institute, Old Road Campus, University of Oxford, Oxford, OX3 7LF, UK.

As improved healthcare leads to older populations, it is increasingly common to find individuals with multiple co-existing diseases (multimorbidity). This article explores whether age and established risk factors are sufficient to explain the incidence rates of multiple, possibly coexisting diseases. By accounting for the limited age-range in UK Biobank data, previous work demonstrated that a Weibull model could accurately describe the incidence of ~ 60% of the most common primary hospital diagnoses of diseases, and those fits are used here to predict the age-dependent incidence of disease with adjustment for established risk factors. A “Poisson binomial” model is combined with these to predict the total number of occurrences of each disease in the UK Biobank cohort that would be expected without pre-existing (prior) disease. The total observed numbers of diseases in men and women (including those from individuals with pre-existing diseases and multimorbidity), were found to be approximately 1.5 times greater than that predicted for individuals without prior disease, and could not be explained by natural statistical variation. The multiple of 1.5 was sufficiently consistent across different diseases to prevent a meaningful classification of disease types with this factor, although there were differences for sub-groups such as smokers with high body mass index, in which there tended to be increases to both the underlying disease risk and to subsequent rates of disease-driven ill health. The study found that for 123 disease in men and 99 diseases in women, age and established risk factors could not explain the incidence rates, that were greater when there was prior disease. It also suggests that empirical modelling might allow reliable predictions of primary causes of hospital admissions, helping to facilitate the planning of future healthcare needs.

Multi-morbidity is increasingly common in developed economies such as the United Kingdom. It is widely believed to involve clusters of diseases and that disease risks are modified by underlying conditions [1–4]. Despite considerable work to characterise multimorbidity in terms of clusters of diseases (e.g. see the recent review [4]), there have been comparatively few recent studies to determine whether pre-existing conditions can modify age-dependent incidence rates of disease [2, 5–11]. Here we restrict attention to primary hospital diagnoses of common diseases, and consider whether the number of disease types that occur in individuals are consistent with expectations based on age and established risk factors. We use results from a recent study [12] that used a Weibull distribution to explore whether the age-related incidence of diseases in UK Biobank [13] are consistent with multistage disease processes [12, 14–16]. Of the 800 common diseases that were considered, 450 were consistent with the model, and there were sufficient cases for 172 (or 156) to be modelled with adjustment for 7 (or 9) established risk factors in men (or women) [12]. By combining these results with a “Poisson Binomial” distribution [17, 18], this article assesses whether differences in disease incidences are different to that expected from natural statistical variation. An outline of the analysis is in figure 1. The statistical method is described next, the results are presented and discussed in the following two sections, and we end with conclusions that summarise the main results.

Independent disease incidence rates

Statistical background. We test the null hypothesis that the incidence rate of each disease type, is independent of the presence of other underlying diseases. Using incidence rates of each disease type, that are estimated for a scenario that approximates no underlying prior disease, we calculate the probability of observing $N_j$ cases of disease type $j$ in the UK Biobank cohort when there is no prior disease. Let $S_{ij}(t)$ be the probability of person $i$ surviving disease $j$ until age $t$, then $p_{ij} = S_{ij}(t_i^{\text{start}}) - S_{ij}(t_i^{\text{end}})$ is the probability of person $i$ first experiencing disease $j$ between the ages $t_i^{\text{start}}$ and $t_i^{\text{end}}$ during which they were in the study. Let $X_{ij} = 1$ if individual $i$ has disease $j$, and zero otherwise. Then if disease risks are independent, the probability of $i = 1$ to $i = n$ individuals observing the set of diseases $X_{ij}$, is

$$P(X_{1j} = x_{1j}, ..., X_{nj} = x_{nj}) = \prod_i p_{ij}^{x_{ij}} (1 - p_{ij})^{1-x_{ij}} \quad (1)$$

The probability of observing $N_j = \sum_i X_{ij}$, where $N_j$ is the number of individuals who experience disease of type $j$, is
Figure 1: The table summarises the data and analysis methods used (top to bottom).

![Table with data and analysis methods](image)

given by the “Poisson Binomial” distribution,

$$P(\sum_{i} X_{ij} = N_j) = \sum_{\{x_{ij} = 0, 1\}} \delta(N_j - \sum_{i} x_{ij}) \Pi_{i} p_{ij}^{x_{ij}} (1 - p_{ij})^{1-x_{ij}}$$  \hspace{1cm} (2)

where $\delta(s)$ is the Dirac delta function that equals 1 when $s = 0$ but is zero otherwise, and the sum is over all values $x_{ij} = 0$ and 1, for $i = 1..n$. The mean and variance of the distribution are,

$$E[N_j\{p_{ij}\}] = \sum p_{ij}$$  \hspace{1cm} (3)

and,

$$Var[N_j\{p_{ij}\}] = \sum p_{ij} (1 - p_{ij})$$  \hspace{1cm} (4)

Simple derivations of these results are given in the Materials and Methods. For $p_{ij} \ll 1$, as is the case for most of the diseases in most individuals considered here [12, 19, 20], then the distribution will approximate a Poisson distribution, which provides another reason why the Poisson distribution is common - it approximates the distribution for the number of events arising from rare independent processes.

For each disease $j$ and individual $i$, the multivariate $\delta$-method [21] allows variances $\sigma_{ij}$ to be calculated for the probability $\hat{p}_{ij}$ of disease $j$ during the time observed, with $\hat{p}_{ij} \sim N(p_{ij}, \sigma_{ij}^{2})$. The law of total variance states that,

$$Var[N_j]\{p_{ij}\} = E[Var(N_j\{p_{ij}\})] + Var[E(N_j\{p_{ij}\})]$$  \hspace{1cm} (5)

that can be evaluated using Eqs. 3 and 4. Surprisingly perhaps, the variances $\sigma_{ij}^{2}$ that arise from integrals over $p_{ij}^{2}$ cancel and do not appear in the result, that is,

$$Var[N_j] = \sum p_{ij} (1 - p_{ij})$$  \hspace{1cm} (6)

and can be estimated by “plugging in” the maximum likelihood estimates $\{\hat{p}_{ij}\}$ for $\{p_{ij}\}$. In practice the estimated variances $\{\hat{\sigma}_{ij}^{2}\}$ were important for “quality control”, allowing the identification of poor quality estimates that are too imprecise, or unlikely to satisfy the assumptions needed for application of the $\delta$-method. Specifically, we excluded diseases with estimates for parameters $x = k$ or $x = L$ where $s.e.(x)/x > 0.5$, or where the $\delta$-method failed to give a numeric estimate, and confirmed that the remaining estimates had $\sum_{i} \sigma_{ij}^{2}/N_j < 0.05$. The negative correlation between parameters $k$ and $L$ led to smaller variances for estimates than might have been expected based on estimates for the variances of $k$ and $L$ individually.

**Data analysis.** Diseases included in the study were defined as a collection of one or more 3- and 4-digit disease codes from the International Classification of Diseases Version 10 (ICD-10) [22], that were selected by three epidemiologists with
individual backgrounds in pathology, general practice, and statistics, based on a set of predetermined criteria as detailed previously [23]. Diagnoses used primary cause of admission in hospital episode statistics, that ensured that the diseases had passed a threshold of severity prior to diagnosis. The study period for an individual was between joining the study and 31st January 2020, after which admission rates will be influenced by the Covid pandemic. Individuals were excluded if they had a prior diagnosis of a cancer other than non-melanoma skin cancer before the study started.

The study considered all primary diagnoses within the study period (“all cases”), and similar to previous work [12, 19, 23], the first primary diagnosis that an individual receives in each ICD-10 chapter (“FIC”). By considering only the first diagnosis in each ICD-10 chapter, the intention was to minimise confounding by prior disease, while maximising the number of cases that are included in the study.

A previous study of UK Biobank data identified 450 diseases whose FIC incidence rates could be modelled by a Weibull distribution [12],

\[ S(t) = \exp \left( -e^{x^T \beta} \left( \frac{t}{L} \right)^k \right) \]

where \( t \) is age, \( k \) and \( L \) are parameters, and \( x, \beta \) are vectors of covariates and parameters. Data were excluded if there was either a report of the disease or a cancer other than non-melanoma skin cancer before the study’s start. Maximum likelihood estimates and their covariances were calculated by left-truncating at the age when participants joined the study, taking age of event as the age of disease-onset, and right-censoring if there is cancer, death or the study ends before disease onset. Adjustment was for the established risk factors of smoking (never, previous, current), diabetes (yes, no), alcohol (rarely - less than 3 times per month, sometimes - more than 3 times per month but less than 3 times per week, regularly - more than 3 times per week), deprivation tertile (min, mid, max), education (degree level, post-16, to age 16), and sex-specific tertiles of height and bmi (min, mid, max). Baseline was never smoked, no diabetes, sometimes drink, min deprivation, degree-level education, middle BMI, minimum height and deprivation tertiles, and women with no HRT use or children.

For each disease whose age-dependent incidence could be modelled sufficiently well, the observed number of cases in the UK Biobank cohort was plotted versus the number of predicted cases if there is no confounding by prior disease (Figure 2). The results for men and women were plotted separately, and excluded diseases whose predicted case numbers differed by more than 4 standard deviations (sd) from the observed numbers of first diseases in each ICD-10 chapter (with minimal confounding by prior disease). The (lenient) threshold of 4 sd was intended to include as many diseases as possible, while limiting the risk of outliers from substantially influencing subsequent results. In practice, for most of the diseases, 4 sd was found to be much less than the differences between the number of observed and predicted cases.

Confidence intervals are reported as \( \pm 1.96 \sigma \), where \( \sigma \) is the estimated standard deviations of the maximum likelihood estimate (MLE). For an MLE \( \hat{\beta}_i, \hat{\beta}_j \sim N(\beta_i, \sigma_i^2) \), and tests for equality of MLEs use \( (\hat{\beta}_i - \hat{\beta}_j) \sim N(0, \sigma_i^2 + \sigma_j^2) \), so statistically significant differences at the 0.05 level will have \( |\hat{\beta}_i - \hat{\beta}_j| > 1.96\sqrt{\sigma_i^2 + \sigma_j^2} \).

Results

Figure 2 plots the number of each disease type observed in the UK Biobank cohort, versus the predicted number of cases if there is no confounding by prior disease, with poorly modelled data removed (as described above). A straight-line fit through the origin estimated the number of observed diseases in men to be 1.50 [1.45,1.55] times the expected number of diseases without prior disease, and 1.53 [1.47,1.59] in women (tables 1 and 2). There is no statistically significant difference between the fits for men and women. The R-squared value for both fits was 0.96.

Subgroups that were expected to include the highest and lowest risk groups were also considered (tables 1 and 2), these included: non-smokers in the mid-BMI tertile, smokers in the mid-BMI tertile, non-smokers in the max-BMI tertile, smokers in the max-BMI tertile. Statistically significant differences between estimates for men and women (at the 0.05 level), are only found for the non-smoking, mid-BMI group, that have lower estimates for females. For both men and women there were statistically significant differences between the group including everyone, and the subgroup of smokers in the max-BMI tertile. For men only, there were statistically significant differences between the group including everyone, and non-smokers in the max-BMI tertile. Overall the increase in disease rates above those expected without prior disease, tended to be higher for groups that would already be expected to have higher disease risk (e.g. smokers and, or, the top BMI tertile).

The number of repeat admissions for the same disease in the same individual were considered, to ensure that they
Table 1: Increased incidence associated with prior disease in men. The estimated slope (Coef), its confidence intervals (C.I.), and $R^2$-squared coefficients for Everyone (Figure 2), and sub-groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>Coef</th>
<th>C.I.</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone</td>
<td>1.50</td>
<td>[1.45, 1.56]</td>
<td>0.96</td>
</tr>
<tr>
<td>Non-smoker, mid-BMI</td>
<td>1.52</td>
<td>[1.46, 1.58]</td>
<td>0.95</td>
</tr>
<tr>
<td>Smoker, mid-BMI</td>
<td>1.56</td>
<td>[1.48, 1.63]</td>
<td>0.93</td>
</tr>
<tr>
<td>Non-smoker, max-BMI</td>
<td>1.69</td>
<td>[1.62, 1.76]</td>
<td>0.95</td>
</tr>
<tr>
<td>Smoker, max-BMI</td>
<td>1.87</td>
<td>[1.77, 1.98]</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Table 1: Increased incidence associated with prior disease in men. The estimated slope (Coef), its confidence intervals (C.I.), and $R^2$-squared coefficients for Everyone (Figure 2), and sub-groups.

Discussion

It had originally been expected that prior disease would increase disease rates, and that these increases would vary substantially between disease types. This would have allowed diseases to be characterised and classified by their sensitivity to prior disease. Instead the increases in disease risk were similar for most of the diseases considered, and between both men and women. This raises the possibility of using empirical modelling to link estimated disease rates without prior disease that are estimated by conventional epidemiological studies, to those that are observed in practice. Furthermore, in contrast to methods that use present trends to predict future demand [6], the (risk-factor adjusted) models here can be used in counterfactual

There were however, statistically significant differences between the increased risks in subgroups such as smokers with high BMI, and the cause of these differences are unclear. It could be that the risk factors are increasing both the underlying disease risk (without prior disease), and the influence of prior disease on your subsequent disease risk. Alternately, the increase in underlying disease risk could lead to a greater number of co-existing disease conditions, that together increase disease risk more than for someone with fewer prior diseases. The particular diseases and mechanisms by which prior disease modifies future disease risk need to be identified and understood. Nonetheless, figure 2 suggests that your overall disease risk is driven by your underlying (disease-free) disease risk, and reducing well-understood disease risks will reduce your overall expected burden of disease in old age.

**Limitations**

This study considered data in UK Biobank, and approximately 60% of the 400 most common diseases whose incidence rates could be modelled accurately with a Weibull distribution [12]. Of these, more stringent fitting requirements detailed in the Methods reduced the number of diseases considered to 123 in men and 99 in women (222 in total). Future work may be able to improve the modelling of disease onset rates, that would allow a more comprehensive study. In addition, the UK Biobank cohort is known to be a poor representation of the UK population as a whole, and it is possible that results might differ for a different cohort. For these reasons, the estimates in figure 2 and tables 1, 2, are likely to be modified for analyses with different cohorts and different collections of diseases. Studies in different cohorts, and with different periods of follow-up, will be needed to explore the generality of these results.

The main purpose of this article was to explore how expected disease rates due to established risk factors and old age, are modified by prior disease. Future work will be needed to establish the diseases and disease-mechanisms that are responsible for modifying subsequent disease risk. Drug use and poly-pharmacy are also likely to be important. It is certain that some medications, such as statins, can modify risks of diseases other than those that they are primarily intended to treat.

**Conclusions**

By combining a Poisson-Binomial distribution with a Weibull model for the age-related incidence of disease, this study considered how much disease risk in the UK Biobank cohort could be explained by age and established risk factors. For the 222 diseases considered, it was found that disease rates were much greater than would be expected without prior disease, or than could be explained by natural statistical variation. The increases in risk were comparatively uniform across the diseases considered, and between men and women. This prevented a meaningful classification of diseases using this factor, but suggests that total disease incidence is approximately proportional to the underlying disease-free disease risks. Specifically, for the diseases in this study, disease rates were increased by approximately 1.5 times those expected of individuals without prior disease, but could be higher in subgroups such as smokers with high BMI. In other words, for the 222 diseases considered, approximately two thirds of primary hospital admissions in the UK Biobank cohort were as expected based on age and established risk factors, with the additional third presumed to be due to prior diseases leading to increased disease risk.

More generally, if the overall disease risks were (approximately) proportional to disease risks without prior disease, there would be several important implications. Firstly, avoiding known risk factors for disease would also reduce the risk, or delay the onset, of multiple diseases in old age. Secondly, it would reaffirm the value of conventional epidemiological studies of disease risk that deliberately avoid potential confounding by prior disease. Thirdly, because the (risk-factor adjusted) model

---

<table>
<thead>
<tr>
<th>Group</th>
<th>Coef</th>
<th>C.I.</th>
<th>R-squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Everyone</td>
<td>1.53</td>
<td>[1.47,1.6]</td>
<td>0.96</td>
</tr>
<tr>
<td>Non-smoker, mid-BMI</td>
<td>1.42</td>
<td>[1.36,1.48]</td>
<td>0.96</td>
</tr>
<tr>
<td>Smoker, mid-BMI</td>
<td>1.60</td>
<td>[1.51,1.69]</td>
<td>0.92</td>
</tr>
<tr>
<td>Non-smoker, max-BMI</td>
<td>1.61</td>
<td>[1.52,1.70]</td>
<td>0.94</td>
</tr>
<tr>
<td>Smoker, max-BMI</td>
<td>1.74</td>
<td>[1.61,1.87]</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Table 2: Increased incidence associated with prior disease in women. The estimated slope (Coef), its confidence intervals (C.I.), and $R$-squared coefficients for Everyone (Figure 2), and sub-groups.
can extrapolate beyond the end of the study period, and the confidence intervals in figure 2 are all comparatively narrow, reliable empirical modelling of future disease rates might be possible. This would be helpful for future healthcare planning.

Further studies are needed to increase the breadth of diseases that can be modelled, to explore the generality of the results in other cohorts, and to test the model’s predictive performance. Despite the present limitations, the results provide the first quantitative characterisation of how prior disease modifies the incidence rates of a wide range of disease types, and a methodology that can be used or developed in future studies.

The Poisson-Binomial model

Simple derivations of the key properties of the “Poisson Binomial” distribution are outlined below, with comprehensive proofs given elsewhere [17, 18].

As discussed in the main text, the probability of observing \( N_j = \sum_i N_{ij} \), where \( N_{ij} \) is the number of different diseases observed in individual \( i \), is given by the “Poisson Binomial” distribution,

\[
P(\sum_i X_{ij} = N) = \sum_{\{x_{ij}=0,1\}} \delta \left( N - \sum_i x_{ij} \right) \Pi_i p_i^{x_{ij}} (1 - p_i)^{1-x_{ij}}
\]  

(8)

where \( \delta(s) \) is the Dirac delta function that equals 1 when \( s = 0 \) but is zero otherwise, and the sum is over all \( x_{ij} \) for \( i = 1..n \). The generating function for 8 is,

\[
G = \sum_{k=0}^{n} e^{ks} P(\sum_i X_{ij} = k) = \sum_{k=0}^{n} \sum_{\{x_{ij}=0,1\}} e^{ks} \delta \left( k - \sum_i x_{ij} \right) \Pi_i p_i^{x_{ij}} (1 - p_i)^{1-x_{ij}} = \sum_{\{x_{ij}=0,1\}} e^{s \sum_i x_{ij} } \Pi_i p_i^{x_{ij}} (1 - p_i)^{1-x_{ij}} = \Pi_i (e^{s p_i} + 1 - p_i)
\]

(9)

where Eq. 8 was included in the second line, the Dirac delta-function led to \( e^{ks} \) being replaced by \( e^{s \sum_i x_{ij}} \) when summing over \( k \) in the third line, the fourth line includes \( e^{s \sum_i x_{ij}} \) within the product over \( i \) and \( j \), and the final line has summed over each \( x_{ij} \) taking values of 0 and 1. Moments of Eq. 8 can be obtained by taking derivatives of Eq. 9 with respect to \( s \), and then evaluating them at \( s = 0 \), with for example,

\[
E[N_j] = \left. \frac{\partial G}{\partial s} \right|_{s=0} = \sum_i p_i e^{s \Pi_{q \neq i} (e^{s p_{qj}} + 1 - p_{qj})} \big|_{s=0} = \sum_{ij} p_{ij} \Pi_{q \neq i} (p_{qj} + 1 - p_{qj})
\]

(10)

Similarly the second moment is,

\[
E[N_j^2] = \left. \frac{\partial^2 G}{\partial s^2} \right|_{s=0} = \sum_i p_i e^{s \Pi_{q \neq i} (e^{s p_{qj}} + 1 - p_{qj})} \big|_{s=0} = \left. \frac{\partial}{\partial s} \sum_i p_i e^{s \Pi_{q \neq i} (e^{s p_{qj}} + 1 - p_{qj})} \big|_{s=0} \right. \]

content...

(11)

where in the final line the first term arises from the derivative of the denominator of the first term in line 3, the second term from the derivative of the numerator of the first term in line 3, and the final term is from the derivative of the product in line 3 similarly to evaluating \( E[N_j] \). Combining Eq. 10 with Eq. 11, the variance can be evaluated as,

\[
E[N_j^2] - E[N_j]^2 = \sum_i p_{ij} (1 - p_{ij})
\]

(12)

\( E[N_j] \) and \( Var[N_j] \) can alternatively be written as \( E[N_j] = n \hat{p} \) and \( Var[N_j] = n Var(\hat{p}) \), where the bars denote averages. If \( p_{ij} \ll 1 \), then the variance Eq. 12 has \( Var[N_j] \simeq \sum_i p_{ij} = E[N_j] \), as would be the case for a Poisson distribution with
rate $\lambda_j = \sum_i p_{ij}$ and $P(N_j) = \lambda_j^{N_j} e^{-\lambda_j} / N_j!$. In fact, when $p_{ij} \ll 1$, the Poisson-Binomial distribution tends to the Poisson distribution [17], as can be seen from the moment generation function Eq. 9, that approximates the Possion distribution for $p_{ij} \to 0$. Using Eq. 9 and expanding in terms of $p_{ij}$,

$$G = \Pi_i \left(1 + p_{ij}(e^s - 1)\right)$$

$$= \exp \left\{\sum_i \log \left(1 + p_{ij}(e^s - 1)\right)\right\}$$

$$= \exp \left\{\sum_i p_{ij}(e^s - 1) + O(p_{ij}^2)\right\}$$

$$= \left((e^s - 1)\lambda_j + O(p^2_{ij})\right)$$

(13)

where $\lambda_j = \sum_i p_{ij}$, and the generating function for a Poisson distribution is $\exp \left((e^s - 1)\lambda_j\right)$. The above argument suggests that provided $p_{ij} \ll 1$, then a Poisson distribution can model the number of events due to $n$ independent processes that each have different but low probabilities $p_{ij}$ of occurring.

Code and data availability

UK Biobank data can be accessed by application through www.ukbiobank.ac.uk. UK Biobank has approval by the Research Ethics Committee (REC) under approval number 16/NW/0274. UK Biobank obtained participant’s consent for the data to be used for health-related research, and all methods were performed in accordance with the relevant guidelines and regulations. R code used to produce figures from summary data will be made available from https://osf.io. R packages used in this study include survival[27], grr[28], data.table[29], and maxLik[30].

Acknowledgments

This research has been conducted using data from UK Biobank, a major biomedical database, under application number 42583. This research was supported by an intermediate research fellowship from the Nuffield Department of Population Health (NDPH), University of Oxford.

References


Aleissa, Faisal Alhazzani, May Moyano, Luis G. Pinhanez, Claudio Gonzalez, Marta C. Gonzalez, Marta C./E-9456-2011 Gonzalez, Marta C./0000-0002-8482-0318 1553-7358.


