Supplementary

Overview of modelling approach

In a nutshell, the modelling approach includes two parts. Firstly, four possible health states for the progression of PCOS were defined and the transition probabilities between these states were estimated. The transition probabilities between the disease states (e.g. from probable to diagnosed PCOS) were based on a Poisson regression model (which is often used for modelling the count data, e.g. the number of events over a period of follow-up period). While the other transition probabilities (e.g. mortality rates) were based on UK census data and life table and some other empirical studies.

As a second step, we used Markov model to simulate the population dynamics of PCOS based on a virtual cohort, whose size was estimated from the census data and prevalence rates in our previous database study (Ding et al., 2016) over a follow-up period of 25 years. For example, the number of PCOS patients aged 15-19 was calculated by the multiplication of the total number of women aged 15-19 on the census data and the prevalence rates of PCOS for this age group. The simulations were performed to allow us to look at the change in the number of individuals ending up in each state with its associated proportions. We were most concerned about the proportion of individuals who develop diabetes over the follow-up period because they would incur significant amount of healthcare costs to the National Health Service in the UK and have substantial reduction in quality of life. Therefore, we recorded the economic and quality of life outcomes associated with each state at each time point (with discounting applied).

Poisson regression model using Bayesian approach

The model can be expressed as a Poisson log-linear model of the following form:

$$d_i \sim Poisson(\lambda_i)$$

$$log(\lambda_i) = \beta_0 + \boldsymbol{\beta X_i} + log(t_i)$$

where d_i is the event indicator representing a Poisson process for individual patient during the entire follow-up period. The value for d_i is either 1 (event) or 0 (censored); β_0 is the baseline hazard rate and $\boldsymbol{\beta}$ is the incremental effects of other variables on the baseline rate for the relevant covariates considered, as collected in the matrix $\boldsymbol{X}_i$. In our case, we consider age groups and specifically set the group 40-44 as the reference. The coefficients $\boldsymbol{\beta}$ represent the log incidence rate ratio of the other age groups compared to the baseline. The variable t_i is the follow-up time for i-th individual and was included as a log offset in the linear predictor. This model was applied to estimate the incidence rates ($\lambda_{12}, \lambda_{23}, \lambda_{13}$). These rates were then converted into transition probabilities (p_{12}, p_{23} and p_{13} as shown in Figure 1 in the main text) across the disease states using the following formula (Gidwani, 2014):

$$p_{rs} = 1 - e^{-\lambda_{rs}} \quad r, s = 1,2,3.$$

The underlying assumption is that the incidence rate is constant for a given age group over a Markov cycle of 1 year. We considered this reasonable as both PCOS and diabetes are chronic diseases and we do not expect the population-level incidence rate to change drastically over a short period of time such as 1 year.

For the incidence rates between State 1 and State 2 (λ_{12}) and between State 1 and State 3 (λ_{13}), a minimally informative prior was used for the baseline incidence rate: $\beta_0 \sim Normal(0,100)$. However, for the incidence rates between State 2 and State 3 (λ_{23}), different versions of prior were tested, based on the information from published studies. We considered three scenarios.

Base case scenario: minimally informative prior for baseline incidence rate and incidence rate ratios in comparison with the other age categories:

$$\beta_0, \beta_1, \beta_2, \dots, \beta_5 \sim Normal(0,100).$$

As all the parameters are on log scale, the prior is indeed vague (because the prior distributions have standard deviation of 10 on the log scale), so as not to exert undue influence on the results. Therefore, in this case, the results would be driven by the data only. In this model specification, we are not including other more or less formal knowledge we may have on the likely range of the rates in our analysis. This is particularly relevant if we consider that the data at hand may be characterised by some bias.

Sensitivity analysis 1: informative prior for baseline incidence rate and incidence ratios were included based on external evidence from published studies. In this scenario, we referred to the incidence of type 2 diabetes in the general population by age and the relative risk of diabetes comparing PCOS patients and the general population to act as prior distributions. Firstly, we considered the incidence rate of type 2 diabetes in the general population aged 40-44; we assumed this to follow a Normal distribution with a mean of 3 per 1000 person-year (PY) with some variability as suggested by Sharma et al. (Sharma et al., 2016). Here, we used β_{pop} to denote this quantity and assumed the prior distribution below (the mean of 3 per 1000 PY was converted on log scale, which results in the value -5.809):

$$\beta_{pop} \sim Normal(-5.809, 0.01).$$

We then defined the relative risk of type 2 diabetes comparing PCOS patients and the general population, following the assumption that PCOS patients are 3 times more likely to develop diabetes (Morgan 2012), which corresponds to the following distribution:

$$\rho \sim Normal(0.81, 0.001)$$

for the parameter ρ, representing the relative risk. Therefore, the baseline incidence rate of women with PCOS aged 40-44 can be computed as:

$$\beta_0 = \beta_{pop} \times \rho.$$

We further defined the relative risk of diabetes comparing the rest 5 age groups with the reference group (represented by $\gamma_1, \gamma_2, \dots, \gamma_5$). Therefore, the incidence rate of the rest age groups on a log scale ($\beta_1, \beta_2, \dots, \beta_5$) can be calculated using the following formula:

$$\beta_i = \gamma_i \times \beta_0 \, , i = 1,2, \dots ,5.$$

Finally we defined the priors for $\gamma_1, \gamma_2, \dots, \gamma_5$ to encode the assumption that, compared with the reference group, the relative risk of diabetes for women aged 15-19, 20-29 and 30-39 are 0.09, 0.37 and 0.63, respectively (Sharma et al., 2016) (M. Sharma et al. 2016). Here we made the assumption that the risk of diabetes in women with PCOS across different age groups follows the same pattern as that for the general population.

$$\gamma_1 \sim Normal(0.504, 0.001)$$

$$\gamma_2, \gamma_3 \sim Normal(0.203, 0.001)$$

$$\gamma_4, \gamma_5 \sim Normal(0.096, 0.001)$$

Sensitivity analysis 2: similarly, the prior distributions in this model were informed by external evidence but in a slightly different way. We identified a case-control matched study using the General Practice Research Data (GPRD) in the UK to examine the incidence rate of type 2 diabetes in PCOS population from 1990 to 2010 (Morgan et al., 2012). The structure of GPRD is similar to THIN and therefore, results from this study were considered to be relevant and appropriate to inform our

model. The overall rate of diabetes (5.7 per 1000 PY) in this GPRD study, denoted as β_{prev} (on log scale), was used to obtain the baseline rate for the reference group and this was achieved by proportional weighting:

$$\exp(\beta_0 + \beta_1) \times w_1 + \exp(\beta_0 + \beta_2) \times w_2 + \cdots + \exp(\beta_0 + \beta_5) \times w_5 + \exp(\beta_0) \times w_6 = \exp(\beta_{prev})$$

where β_{prev} represents the overall rate on log scale estimated from the GPRD study; β_0 is the baseline rate for the reference age group and $\beta_1, \beta_2, \ldots, \beta_5$ were defined as the relative risk of diabetes comparing the rest age groups with the reference age group; $w_1, w_2 \ldots, w_6$ are the proportion of women in each age group estimated from our THIN cohort.

By rearranging the formula, we obtained an equation that can be used to compute the baseline rate for the reference group:

$$\beta_0 = \beta_{prev} - \log\left\{\sum_{n=1}^{5}[w_i \times \exp(\beta_i)] + w_6\right\}.$$

The assumption here is that the age distribution of PCOS cohort is similar between our study and the GPRD study. This was considered reasonable because the GPRD also collects primary care data from practices across UK and the structure of the database is largely similar to THIN. The following prior distribution was included for β_{prev}, which corresponds to an incidence rate of 5.7 per 1000 PY with limited variability (i.e. a variance of 0.01):

$$\beta_{prev} \sim Normal(-5.167, 0.01).$$

We further defined a prior for the relative risk of diabetes across age groups and the distributions below indicate that the relative risk of diabetes is 0.09, 0.37 and 0.63 for women aged 15-19, 20-29 and 30-39, respectively, compared with the reference group:

$$\beta_1 \sim Normal(-2.408, 0.001)$$

$$\beta_2, \beta_3 \sim Normal(-0.994, 0.001)$$

$$\beta_4, \beta_5 \sim Normal(-0.462, 0.001)$$

It should be noted that the small variances of the prior distributions we included for Scenario 2 and 3 are consistent with the data provided in the published studies. Moreover, we have used forward sampling to graphically verify the mean and the associated 95% CI for all the prior distributions to ensure that they are in line with the published studies. Given the large sample size in both our study and the previous study (e.g. population study), we would expect low uncertainties in our model parameters.

Model convergence

We ran this model with two Markov chains starting at arbitrary values for convergence purpose. A total of 50,000 simulations per chain were generated and the first 5,000 in burn-in period were discarded. The number of thinning was set to be 90 to reduce the autocorrelation so the final number of simulations saved are 1000. The model convergence was assessed based on the Gelman-Rubin diagnostic statistics, which were provided for all parameters in the model in the plot below. For the Gelman-Rubin statistic, there is a cut-off point of 1.1, values below which indicate convergence of the MCMC procedure to the target posterior distributions.

Rationales for Markov model

The baseline year was set to be 2014. Therefore, the number of patients in each state at the baseline year was estimated using the census data in mid-2014 multiplied by the prevalence rates in 2014 (Ding et al., 2016).

The initial distribution of cases over states was assumed to be consistent with that from the database we used for this analysis (i.e. THIN, as introduced in the main text). For example, 18.2% of the diagnosed cases had a prior diagnosis of diabetes and these women would start in State 3 (PCOS with diabetes) rather than State 2 (Diagnosed PCOS).

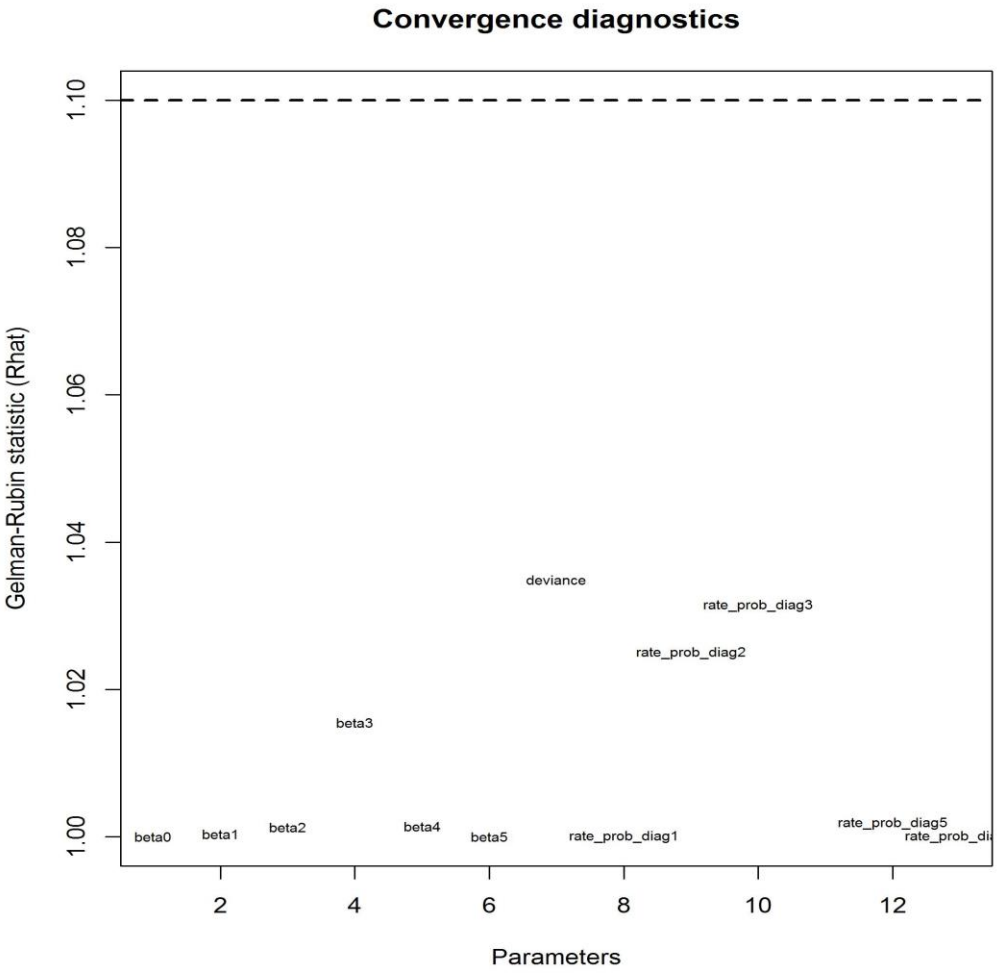
The follow-up period was set to be 25 years. The rationale is that this time period was considered to be relatively long to examine the transition from PCOS to diabetes for younger women in our cohort. There is evidence that the incidence of type 2 diabetes increases dramatically for women aged 50-70 (Sharma et al., 2016) and after 25 years of follow-up, the majority of our study population was expected to be within this age range.

As the rates estimated from THIN data are only for reproductive-aged women, we assumed that patients who exceed the age of 45 during the follow-up develop diabetes at the same rate as those aged 40-45. Probable cases were assumed to be no longer able to receive a confirmed diagnosis after the age of 45 because the major symptom of PCOS (i.e. menses) disappears after menopause.

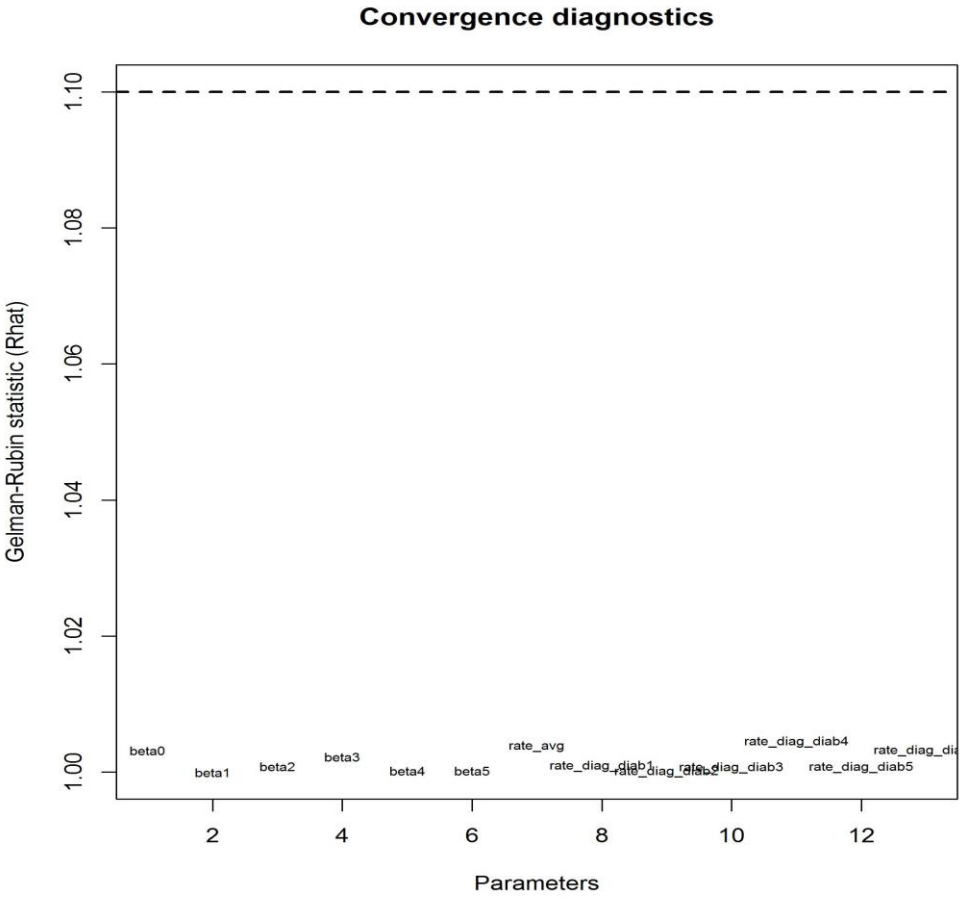
Two scenarios were considered. In the first scenario, a closed cohort with PCOS population aged 15-44 was simulated. In the second scenario, we considered an open cohort model with females aged 1-14 in 2014 gradually entering the study population in the follow-up period. The total number of cases aged 1-14 was calculated as the sum of probable and diagnosed cases and the prevalence used for calculation is consistent with that for the age group of 15-19. The assumption here is that women aged 1-14 are diagnosed at same rates as the youngest age group once they reach the age of 15 and similarly, the distribution of patients over states after they enter into the “observable" period is consistent with that from THIN as explained previously.

Note that sensitivity analysis was performed to assess the likely impact of incidence rates estimated based on different model assumptions (i.e. base case scenario, Sensitivity analysis 1 and 2) on the outcome (i.e. proportion of patients who develop diabetes by the end of follow-up).

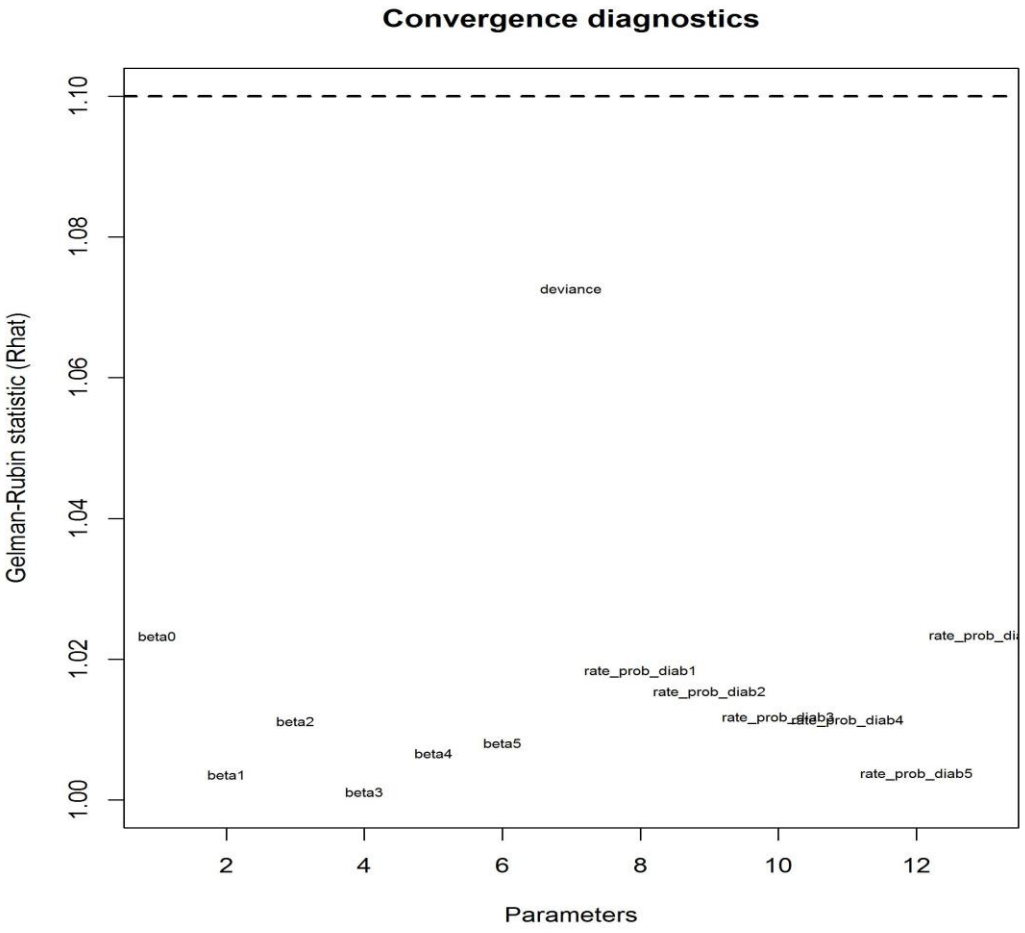
Supplementary Figure I Convergence assessment for model estimating the incidence rates from probable PCOS (State 1) to diagnosed PCOS (State 2). The convergence diagnostics (Gelman-Rubin statistic) for all the parameters are provided: beta_i and rate_prob_diag_i (i=1,2,…,6) represents the incidence rates on natural and log scale estimated from the model for the six age groups, respectively. Note that the cut-off point of Gelman-Rubin statistic is 1.1 and values below indicate that the model converges.



Supplementary Figure II Convergence assessment for model estimating the incidence rates from diagnosed PCOS (State 2) to diabetes (State 3). The convergence diagnostics (Gelman-Rubin statistic) for all the parameters are provided: beta_i and rate_diag_diab_i (i=1,2,…,6) represents the incidence rates on natural and log scale estimated from the model for the six age groups, respectively. Note that the cut-off point of Gelman-Rubin statistic is 1.1 and values below indicate that the model converges.



Supplementary Figure III Convergence assessment for model estimating the incidence rates from probable PCOS (State 1) to diabetes (State 3). The convergence diagnostics (Gelman-Rubin statistic) for all the parameters are provided: beta_i and rate_prob_diab_i (i=1,2,…,6) represents the incidence rates on natural and log scale estimated from the model for the six age groups, respectively. Note that the cut-off point of Gelman-Rubin statistic is 1.1 and values below indicate that the model converges.



Supplementary Table I: Transition probabilities between states in the Markov model.

Parameter	Definition	Source of data
p_{12}	Transition probabilities from probable PCOS to diagnosed PCOS	Estimated from the PCOS cohort extracted from THIN
p_{13}	Transition probabilities from probable PCOS to diabetes	Estimated from the PCOS cohort extracted from THIN
p_{23}	Transition probabilities from diagnosed PCOS to diabetes	Estimated from the PCOS cohort extracted from THIN with prior informed from published studies
p_{14}, p_{24}	Transition probabilities from probable and diagnosed PCOS to death	Assumption based (PCOS itself does not increase mortality and mortality rate is estimated using rates for the general population), refer to mortality rates in female population by age in the UK
p_{32}	Transition probabilities from diabetes to death	Mortality rates in diabetes population, refer to published studies (Mulnier et al., 2006)

Supplementary Table II. Cumulative proportion of cases receiving relevant prescriptions for PCOS within 1 year after their diagnosis (by case definition).

Treatment	Percentage of cases receiving prescription	
	Probable cases	Diagnosed cases
Combined oral contraceptives	12.8%	17.5%
Progestogen oral contraceptives	4.8%	5.9%
Metformin	1.7%	17.8%
Eflornithine	1.8%	4.6%
Weight loss/control drug	0.63%	2.3%
Acne drugs	24.9%	11.1%

Supplementary Table III Recommended dose and treatment instruction for each drug considered.

Treatment	Dose and instruction
Combined oral contraceptive/Progestogen oral contraceptives	One tablet daily for 21 days and repeat for each menstrual cycle until menopause
Metformin	500-1000mg daily to start with for the first week and then 1000-1500mg daily for the second were and 1500-2000mg daily if tolerated thereafter for 3-6 month
Eflornithine	Apply twice daily and should be discontinued in the absence of improvement after treatment for 4 months.
Weight loss/control drug	Orlistat: 120mg for maximum 3 times daily and continue treatment beyond 12 weeks only if weight loss since start of treatment exceeds 5%
Acne drugs	Most are topical cream or gel including erythromycin, benzoyl peroxide, tretinoin and isotretinoin. Apply 1-2 times daily and review at 8 weeks. Treatment may take up to 6 months or beyond depending on severity.

Reference

Ding, T., Baio, G., Hardiman, P. J., Petersen, I., and Sammon, C.'Diagnosis and management of polycystic ovary syndrome in the uk (2004-2014): a retrospective cohort study. *BMJ open* 2016;**6**:e012461.

Gidwani, R. Deriving Transition Probabilities for Decision Models, 2014. https://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/819-notes.pdf

Morgan, C. L., Jenkins-Jones, S., Currie, C. J., and Rees, D. A. Evaluation of adverse outcome in young women with polycystic ovary syndrome versus matched, reference controls: a retrospective, observational study. *The Journal of Clinical Endocrinology & Metabolism* 2012;**97**:3251-60.

Mulnier HE, Seaman HE, Raleigh VS, Soedamah-Muthu SS, Colhoun HM, Lawrenson RA. Mortality in people with type 2 diabetes in the UK. *Diabetic medicine : a journal of the British Diabetic Association* 2006;**23**:516-21.

Sharma, M., Nazareth, I., and Petersen, I. Trends in incidence, prevalence and prescribing in type 2 diabetes mellitus between 2000 and 2013 in primary care: a retrospective cohort study. *BMJ Open* 2016;**6**:e010210.