

1 **Machine learning framework for predicting the presence of high-risk clonal haematopoiesis using**
2 **complete blood count data: a population-based study of 431,531 UK Biobank participants**

3 William G. Dunn*^{1,2,3} MBChB, Isabella Withnell*⁴ MSc, Muxin Gu^{1,2} PhD, Pedro Quiros⁵ PhD, Sruthi
4 Cheloor Kovilakam^{1,2} PhD, Ludovica Marando⁶ MD, Sean Wen⁷ DPhil, Margarete A Fabre^{2,3,7}
5 MBChB, Irina Mohorianu^{† 1} PhD, Dragana Vuckovic^{† 8} PhD, George S. Vassiliou^{† 1,2,3} MBBS

6

7 *[†] Contributed equally

8

9 1. Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK.

10 2. Department of Haematology, University of Cambridge, Cambridge, UK.

11 3. Department of Haematology, Cambridge University Hospitals NHS Trust, Cambridge, UK.

12 4. Division of Biosciences, University College London, London, UK.

13 5. Department of Biochemistry and Molecular Biology, Instituto Universitario de Oncología
14 (IUOPA), Universidad de Oviedo, 33006, Oviedo, Spain.

15 6. Department of Hematology, Mayo Clinic, Rochester, MI, USA.

16 7. Centre for Genomics Research, Discovery Sciences, Biopharmaceuticals R&D, AstraZeneca,
17 Cambridge, UK.

18 8. Department of Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine,
19 Imperial College London, London, UK.

20

21 **Abstract**

22 **Background**

23 Clonal haematopoiesis (CH), the disproportionate expansion of a haematopoietic stem cell and its
24 progeny, driven by somatic DNA mutations, is a common age-related phenomenon that engenders an
25 increased risk of developing myeloid neoplasms (MN). At present, CH is identified by targeted
26 sequencing of peripheral blood DNA, which is impractical to apply at population scale. The complete
27 blood count (CBC) is an inexpensive, widely used clinical test. Here, we explore whether machine
28 learning (ML) approaches applied to CBC data could predict individuals likely to harbour CH and
29 prioritise them for DNA sequencing.

30

31 **Methods**

32 The UK Biobank was filtered to identify 431,531 participants with paired CBC and whole exome
33 sequencing (WES). Somatic mutations were previously identified from blood WES using Mutect2 to
34 classify individuals with CH driver mutations. Using 18 CBC indices/features and basic demographics
35 (age and sex), we trained a range of tree-based ML classifiers to infer as binary output, the presence/
36 absence of CH.

37

38 **Findings**

39 Using Random Forest (RF) classifiers, we predicted the presence/absence of CH driven by mutations
40 in one of five genes known to confer a high-risk of incident MN (*JAK2*, *CALR*, *SF3B1*, *SRSF2* and
41 *U2AF1*). We subsequently developed a unified, optimised RF classifier for high-risk CH driven by any
42 of these genes and assessed its performance (median AUC 0.85). However, the low prevalence of high-
43 risk CH implies that our model cannot be generalised to population scale without compromising its
44 sensitivity (20.1% using stringent cutoff probability score).

45

46 **Interpretation**

47 We showcase a proof-of-concept that the presence of high-risk CH can be inferred from CBC
48 perturbations using RF classifiers. The future integration of raw blood cell analyser data can help
49 improve the performance of our model and facilitate its application at scale.

50

51 **Funding**

52 Cancer Research UK.

53

54

55

56 **Research in context**

57 **Evidence before this study**

58 We searched PubMed for articles published, in English, between database inception and 5th of June
59 2024, using the terms “clonal hematopoiesis” AND (“machine learning” OR “artificial intelligence”).
60 We additionally searched for the terms “clonal hematopoiesis” AND “complete blood count”. We found
61 18 research articles: one article used ML approaches (XGBoost classifiers) to differentiate clonal
62 haematopoiesis “driver” mutations from “passenger” mutations, but none linked machine learning
63 frameworks to complete blood count data for predicting the presence of clonal haematopoiesis.
64 Progression from clonal haematopoiesis to myeloid neoplasia is known to be associated with several
65 blood count parameters; two recent publications developed clonal haematopoiesis risk stratification
66 tools that incorporated blood count indices in their final risk prediction models (Gu et al. Nature
67 Genetics, Weeks et al. NEJM Evidence). However, we found no study assessing whether blood count
68 indices could be used to infer the presence of clonal haematopoiesis.

69

70 **Added value of this study**

71 Here we show that CH driven by mutations in genes associated with high risk of progression to myeloid
72 neoplasia can be reliably differentiated using ML approaches applied on peripheral blood indices;
73 however, low-risk forms of CH (driven by mutations in the *DNMT3A* or *TET2* genes) cannot be reliably
74 inferred from CBC indices. While optimising the model we identified challenges in upscaling its
75 applicability; we propose that the integration of single-cell resolution “raw” blood analyser data might
76 overcome these issues. Previous efforts to enhance the scalability of CH screening focused on reducing
77 DNA sequencing costs. Here, we provide a proof-of-concept that an extensively used clinical test, the
78 CBC, can, using machine learning approaches, predict individuals more likely to harbour high-risk CH,
79 who should be prioritised for genetic testing.

80

81 **Implications of all the available evidence**

82 Our study proposes a model for predicting high-risk CH mutations by applying a Random Forest
83 classifier on CBC indices; this represents an important step towards scalable screening for identifying
84 individuals at high risk of developing myeloid neoplasia in the future. This is an attractive approach, as
85 it relies solely on a routine, inexpensive test. Despite good sensitivity, the low prevalence of high-risk
86 CH leads to a low positive predictive value that precludes the use of the predictive model as a
87 population-wide pre-screening tool. To overcome this, we propose the future integration of raw blood
88 analyser data into models like ours to improve the performance and scalability of this approach.

89 Introduction

90 Haematopoiesis, the formation of the cellular components of blood, occurs continuously throughout
91 life. At steady state, haematopoiesis generates $4\text{-}5 \times 10^{11}$ cells per day¹⁻⁴ and this vast output is
92 maintained by a small pool of 50,000-200,000 multipotent haematopoietic stem cells (HSCs)⁵ through
93 a cascade of differentiation and proliferation. Somatic mutations accumulate during life, and though
94 most are inconsequential, some can enhance cellular fitness and are positively selected in
95 physiologically normal tissues⁶⁻⁸. Clonal haematopoiesis (CH) is an age-related phenomenon that arises
96 when a HSC acquires a somatic driver mutation (i.e. one that increases its fitness), leading to clonal
97 expansion of the cell and its progeny^{9,10}. Large population-based studies revealed that the most
98 commonly mutated genes in CH are involved in epigenetic regulation (*DNMT3A*, *TET2*, *ASXL1*), signal
99 transduction (*JAK2*, *GNBI*), DNA damage response and apoptosis (*TP53*, *PPM1D*), and splicing
100 (*SF3B1*, *SRSF2*, *U2AF1*)⁹⁻¹⁴. The prevalence of CH increases with advancing age to affect at least 20%
101 of those over 70 years, in whom the phenomenon is almost universally detectable when deep sequencing
102 approaches are employed⁹⁻¹⁴.

103

104 A hallmark of CH is the associated increased risk of incident myeloid neoplasms (MN), a molecularly
105 heterogeneous group of blood cancers that include acute myeloid leukaemia (AML), myelodysplastic
106 syndromes (MDS) and myeloproliferative neoplasms (MPN). The overall rate of progression to MN is
107 low (0.5-1% per annum)⁹, but the risk and nature of malignant progression vary according to the mutant
108 driver gene, the size of the clone, and the selection pressures to which the clone is exposed^{15,16}. Recent
109 advances have facilitated the precise estimation of the risk of progression from CH to MN^{16,17}, such that
110 individuals at high risk can be identified and prioritised for clinical follow-up. CH may precede the
111 development of MN by years^{9-11,15,16,18}, and this provides a window during which high-risk clones could
112 be intercepted and targeted to avert or delay the development of MN.

113

114 A key impediment to prospective myeloid cancer prevention programmes is the lack of a scalable test
115 to identify CH. At present, CH is identified by Next Generation Sequencing (NGS) of blood DNA
116 targeted to a panel of genes recurrently mutated in MN. However, NGS is not performed in routine
117 clinical practice and is impractical and costly to perform at scale. An alternative approach is to leverage
118 low-cost, scalable, routine clinical tests to identify individuals likely to harbour CH who can be
119 prioritised for sequencing. The complete blood count (CBC) is an inexpensive, routine clinical test, and
120 CBC indices such as the red cell distribution width (RDW) and mean cell volume (MCV) are associated
121 with progression from CH to MN¹⁸. We therefore sought to explore whether machine learning (ML)
122 models could predict individuals with CH based on CBC features by analysis of paired CBC and whole
123 exome sequencing (WES) data from 431,531 United Kingdom Biobank (UKB) participants.

124

125 Methods

126 **Study design and participants**

127 We utilised data from the UKB (<https://www.ukbiobank.ac.uk/>), a population-based cohort of 502,536
128 volunteers recruited to the United Kingdom recruited between 2006-2010 and aged between 37 and 73
129 years at recruitment¹⁹. Participants' data was accessed under approved UKB applications number 56844
130 and 69328.

131
132 To derive a dataset for use in our ML pipeline, we excluded UKB participants with any missing CBC
133 variables and those without WES data. Since CH is defined by the presence of a leukaemia-associated
134 somatic driver mutation in an individual without an apparent blood neoplasm, participants with a
135 previous diagnosis of a haematological malignancy were excluded from the final dataset, as were those
136 who developed an incident haematological malignancy within 30 days of recruitment to the UKB. After
137 exclusions, 431,531 participants were retained for downstream analyses.

138

139 **Variable selection**

140 We extracted all CBC variables measured in the UKB ($n=22$), and augmented the feature set with the
141 participants' age and sex. Some CBC variables are closely related or derived from one another; to assess
142 collinearity we computed a pairwise Spearman's rank correlation coefficient (r_s) and excluded variables
143 with a $|r_s| \geq 0.9$. This led us to exclude haematocrit, high light scatter reticulocyte count and the total
144 white blood cell count, whilst retaining their highly correlated counterpart features (haemoglobin
145 concentration, reticulocyte count and neutrophil count, respectively). Nucleated red blood cell count
146 (NRBC) was also excluded as it exhibited near-zero variance (106 unique values, NRBC=0 in 98.9%
147 of UKB participants).

148

149 **Identification of clonal haematopoiesis from whole exome sequencing data**

150 CH was identified from whole exome sequencing (WES) of blood DNA from 431,531 UKB participants
151 as previously described¹⁶ (see Supplement). UKB participants were subsequently labelled as "any-
152 driver-CH" or "no CH" based on the presence or absence of a driver mutation(s) at VAF $\geq 2\%$. For input
153 to gene-specific models of CH, we additionally labelled UKB participants by driver gene (e.g. "*TET2*-
154 CH", "*SRSF2*-CH", etc vs "no CH"). Individuals with ≥ 2 driver mutations were labelled on the gene
155 with the highest VAF.

156

157 **Supervised machine learning model development**

158 Having derived ground truth levels from WES data, ML models were subsequently built for "any-
159 driver-CH" (variant allele frequency, VAF, $\geq 2\%$ with a driver mutation in any CH gene), "large clone
160 any-driver-CH" (as previous but VAF $\geq 10\%$), and each driver gene CH subtype.

161

162 To develop a binary classifier for predicting the presence/absence of CH, we trained and evaluated a
163 selection of tree-based machine learning models: Decision Trees, Random Forests and Extreme
164 Gradient Boosting (XGBoost) Trees. Tree-based approaches were preferred since the set of input
165 features was heterogeneous (continuous and categorical); moreover, these models, augmented with
166 statistical analyses, may also capture the interaction between features. Aside from the assessment of
167 near-zero variance and collinearity, no further pre-processing was applied to the input dataset.

168
169 All 18 CBC parameters were used as features, in addition to basic demographic data (age at sampling
170 and sex). Since the UKB CH dataset was imbalanced, with significantly more controls (no CH) than
171 cases (CH), a random down-sampling was performed to achieve a 1:1 ratio of cases:controls in the input
172 data, to enhance model training and convergence; this down-sampling process was repeated ten times
173 iteratively (Supplementary Figure 1). Subsequently, down-sampled datasets were partitioned on 80:20
174 training:test ratio.

175
176 All models were built using ten repeats of ten-fold cross-validation setups; a grid-search approach was
177 used to tune the relevant hyperparameters (Supplementary Table 1). To avoid technical bias from the
178 down-sampling step, a modified cross-validation was applied, training and evaluating each ML model
179 ten times iteratively, each time using a different random down-sample of the majority (control) class,
180 thereby quantifying the robustness and stability of each model to variation in the subset of control
181 samples or train/test partition (Supplementary Figure 1). Model performance was assessed on the
182 unseen test data, on receiver operating characteristic (ROC) curves and area under the curve (AUC), in
183 addition to sensitivity and specificity.

184
185 From the Random Forests models, we determined variable importance by computing the mean decrease
186 in node impurity from splitting on each feature (measured by Gini index), averaged across all trees and
187 across each of the ten repeats of model-building, using the `importance()` function from the `randomForest`
188 package in R (v4.7.1)²¹. The consistency across top-ranked variables per driver was visualised using
189 quantitative Venn diagrams (upset plots, `ComplexUpset` package) on the top two variables. The feature
190 selection was performed by ranking all n features by importance, in descending order, and iteratively
191 excluding the least informative feature, to determine a minimum set of highly predictive features.

192
193 To assess the scalability of the final model in a “real-life” setting, i.e. with class imbalance (more
194 controls (no CH) than cases (CH)), we added unseen control cases to the test set to match the prevalence
195 of CH cases in the test set to the prevalence of CH cases in the UKB cohort. We examined the trade-off
196 of sensitivity (which is independent of prevalence), positive predictive value (which is dependent on
197 prevalence) and the model prediction score, using this to determine the optimal cut-off score, that
198 minimises the false positives whilst retaining adequate sensitivity.

199

200 All ML models were built using the Caret v6.0.91 package in R v3.6.3²⁰. A full list of packages used is
201 available in Supplementary Methods. All code used to implement our ML framework is publicly
202 available on GitHub: <https://github.com/billydunn/chic>.

203

204 **Results**

205 After excluding those with missing CBC data (n=32,670), missing WES data (n=36,368), or a prevalent
206 diagnosis of a haematological malignancy (n=1840), CH (VAF $\geq 2\%$) was identified in 20,860/431,531
207 (4.8%) UKB participants, of whom 7637/20,860 (36.6%) had large clone CH (VAF $\geq 10\%$; Figure
208 1, Table 1). Using this UKB dataset, we developed a range of tree-based models using our ML
209 framework, which we henceforth refer to as CHIC (Clonal Haematopoiesis Inference from Counts).

Driver	n	Proportion (%)	Male (%)	Large Clone (%)
No driver	410671	95.17	45.82	N/A
Multiple drivers	868	0.2	54.26	70.16
<i>DNMT3A</i> non-R882	8885	2.06	40.89	33.04
<i>TET2</i>	4530	1.05	47.77	36.95
Other driver	1866	0.43	47.27	25.35
<i>DNMT3A</i> R882	1708	0.4	41.74	43.62
<i>ASXL1</i>	1706	0.4	64.54	43.38
<i>PPM1D</i>	720	0.17	56.39	31.81
<i>TP53</i>	402	0.09	54.23	31.59
<i>SRSF2</i>	260	0.06	77.69	61.54
<i>SF3B1</i>	211	0.05	68.25	59.24
<i>GNB1</i>	178	0.04	32.02	76.97
<i>JAK2</i>	167	0.04	50.3	80.24
<i>CALR</i>	104	0.02	60.58	57.69
<i>IDH2</i>	70	0.02	64.29	62.86
<i>U2AF1</i>	53	0.01	75.47	100

210

211 *Table 1: Clonal haematopoiesis proportions in the filtered cohort (n = 431,531).*

212

213 We firstly examined whether CH could be predicted from CBC data in the UKB using models agnostic
214 to underlying driver mutations (henceforth “any-driver-CH”). Using CHIC we generated binary
215 classifiers (CH/no CH) of any-driver CH using tree-models with 18 CBC variables augmented with age
216 and sex as features. Classifiers of any-driver CH were better than random, but with limited performance
217 across all model types (median AUC on unseen test set 0.62, 0.64 and 0.62 for DT, RF and XGB models
218 respectively) (Figure 2A).

219

220 CH is a molecularly heterogeneous entity, and we posited that the nature and strength of the CBC
221 phenotype conferred by a somatic mutation may vary according to the specific driver gene. We trained
222 driver gene-specific binary classifiers (with labels driver gene CH/no CH) using the same input
223 variables as for the any-driver CH models. The most prevalent forms of CH, driven by mutations in
224 *DNMT3A* and *TET2*, were not robustly detectable; this conclusion held for *DNMT3A*-R882 hotspot
225 mutations, which are associated with a slightly higher risk of transformation to AML¹⁶ (median AUC
226 0.60, 0.62 and 0.64 for *DNMT3A*-R882, *DNMT3A*-non R882 and *TET2* RF models respectively) (Figure
227 2B). By contrast, CH driven by lower prevalence but higher risk driver mutations in the genes *JAK2*,
228 *CALR*, *SF3B1*, *SRSF2* and *U2AF1* performed well (median AUC 0.94, 0.91, 0.84, 0.82, 0.84
229 respectively for RF models) (Figure 2B). Since Random Forests (RF) models generally exhibited the
230 best performance across the driver genes (Figure 2B, Supplementary Table 2), we focused on further
231 developing and exploring RF models.

232
233 CH is strongly associated with age, whilst some driver genes exhibit sex bias. To understand the
234 influence of age and sex in the RF models, we trained each set of driver gene-specific RF models in
235 three iterations: i) with age and sex as the only features, ii) with CBC indices as the features, whilst age-
236 and sex-matching cases to controls (to capture the predictive performance of CBC alone), and iii) with
237 age, sex and CBC indices as features, without age- and sex-matching of cases/controls (to capture the
238 predictive performance of both basic demographics and CBC indices). The performance of models
239 trained with only age and sex as features was generally poor (median AUC <0.75 in all cases, Figure
240 2C); an exception was the age/sex-only model of *SRSF2*-CH, in line with the sharp rise in prevalence
241 of *SRSF2*-CH with advancing age and its strong association with male sex¹⁵.

242
243 Classifiers of CH driven by high-risk genes *JAK2*, *CALR*, *SF3B1*, *SRSF2* and *U2AF1* also performed
244 best when using CBC indices as features and age/sex matching cases to controls in the training and test
245 sets. The predictability of the presence of CH driven by mutations in splicing factor genes (*SF3B1*,
246 *SRSF2* and *U2AF1*) was augmented when age and sex were added as features and age/sex matching
247 was omitted. Acknowledging the age and sex predictive power, we added these features to CBC indices
248 in subsequent models.

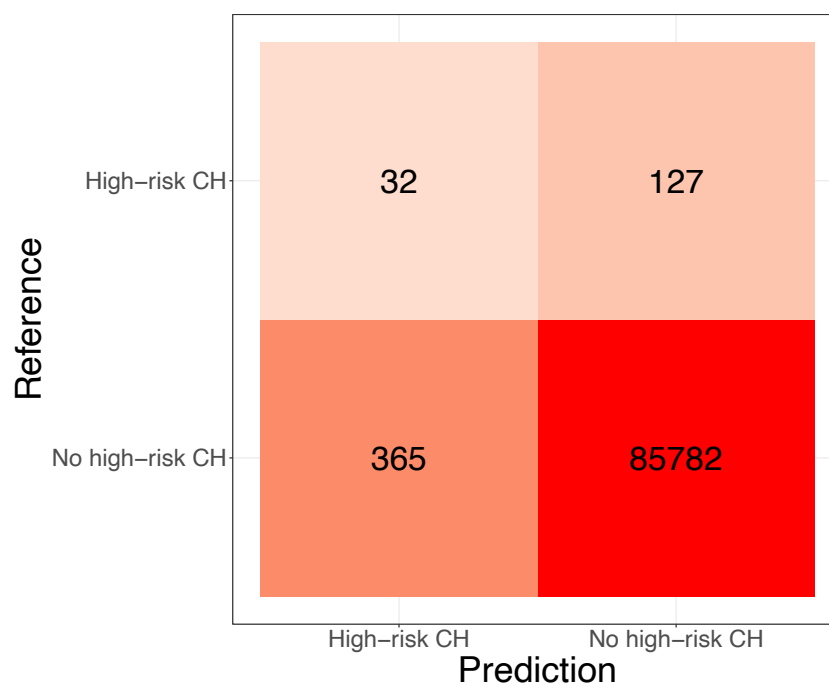
249
250 Since CH with mutations in any of *JAK2*, *CALR*, *SF3B1*, *SRSF2* or *U2AF1* was more predictable from
251 CBC indices and more clinically relevant (associated with high risk of progression to MN), we next
252 combined all predictors into a single binary classifier of “high-risk CH”, to predict the presence/absence
253 of a mutation in any of these five genes (training on input data labelled as “high-risk CH” vs “no high-
254 risk CH”). The resulting median AUC was 0.85 on the unseen test set, Figure 3A); the model also
255 predicted the presence of large (VAF $\geq 10\%$) high-risk clones (median AUC on unseen test set 0.90,
256 Supplementary Figure 2).

257

258 To further refine the classifier of high-risk CH with VAF $\geq 2\%$, we performed iterative feature selection,
259 incrementally excluding the least discriminative feature, to obtain the minimal stable set of highly
260 discriminative features; this demonstrated that our classifier of high-risk CH had undiminished
261 performance using only six features: age at blood sampling, red cell distribution width (RDW), platelet
262 count, platelet distribution width (PDW), platelet crit and mean corpuscular haemoglobin (MCH)
263 (Figure 3B-C). We therefore chose this compact high-risk CH model to explore further, selecting the
264 model that most closely approximated the median AUC across the ten models built using our iterative
265 pipeline.

266

267 Next, we assessed the optimal prediction score cut-off (threshold) for our compact high-risk CH model
268 by examining the trade-off between sensitivity and positive predictive value (PPV) (Figure 3D). In our
269 UKB cohort, high-risk CH was rare (795/431,531 UKB participants, prevalence 0.18%): since the PPV
270 is strongly influenced by the prevalence of positive cases, this necessitated the use of a stringent
271 prediction score cut-off to minimise the number of false positives. To achieve this, we chose a cut-off
272 probability of 0.925, giving a PPV of 8.1% and sensitivity of 20.1% in our unseen test cohort
273 ($n=86,306$), whilst maintaining the specificity and negative predictive value (NPV) of $>99.5\%$ (Table
274 2).



275

276 **Table 2:** Confusion matrix for predictions made by the classifier of high-risk CH on the unseen test set
277 ($n=86,306$), using the previously determined stringent cutoff to determine whether the classifier
278 predicts the positive class (high-risk CH) or the negative class (no high-risk CH).

279

280 A key limitation of the UKB is the low WES coverage, with the driver genes *JAK2*, *SF3B1* and *U2AF1*
281 all having a median coverage of ≤ 31 reads¹⁶, rendering variant calling insensitive to smaller clones. As
282 such, we examined outcomes for the 365 “false positive” cases identified by our high-risk CH classifier,
283 and found that 38/365 (10.4%) developed MN at a median of 5.2 years from sampling. By contrast,
284 only 317/85,782 (0.4%) percent of “true negatives” developed MN. Since CH is the shared precursor
285 of the vast majority of MNs, these observations strongly suggest that the “false positive” individuals
286 had CH below the limit of detection of WES.

287

288 To further explore this hypothesis, we searched for low VAF hotspot mutations amongst 38 individuals
289 who developed MPN, but were not found to have this hotspot mutation by standard variant calling. To
290 do so, we used “pileup” to detect hotspot mutant reads that were filtered out by the stringent criteria of
291 standard calling; this revealed that 13/38 of apparently false positives who developed incident MN had
292 detectable CH mutations by this method, including 11 with driver mutations in *JAK2*, a low coverage
293 gene. This strongly suggests that we underestimated our model performance due to the constraints of
294 WES.

295

296 Further examination of cases identified by CHIC revealed an enrichment in cases with thrombocytosis,
297 suggestive of undiagnosed or unannotated MPN rather than CH (Supplementary Figure 3). Similarly, a
298 few cases had cytopenias that would fall into the diagnostic criteria for CCUS (clonal cytopenia of
299 undetermined significance) or MDS²². To overcome this, we constrained our training/test sets to
300 individuals without cytopenias, thrombocytosis or erythrocytosis, (see Supplementary Methods) and
301 retrained our high-risk classifier. This led to only a minor reduction in performance (median AUC on
302 unseen test set 0.80, Supplementary Figure 4), however, this exacerbated the trade-off between
303 sensitivity and PPV, leading to sensitivity and PPV of only 11.3% and 2.0% respectively at our proposed
304 cutoff probability of 0.875 (Supplementary Figure 4).

305

306 In addition to their use for prediction, CHIC ML models could also uncover novel associations between
307 driver mutations and CBC indices. By evaluating variable importance across all the driver-gene-specific
308 classifiers, and summarising the overlap between the top two features in each model (Figure 4A), we
309 observed known or expected associations: age was highly predictive across models, *JAK2*-CH and
310 *CALR*-CH shared platelet count and platelet crit as important features whilst MCV was predictive of
311 *SF3B1*-CH. Unexpected associations were also revealed: for example, the basophil count was
312 discriminative for predicting the presence *GNBI*-CH only, whilst eosinophil count was discriminative
313 for the presence of *IDH2*-CH. Examining the distribution of each of these CBC variables in the UKB,
314 we found that individuals with *GNBI*-CH had a significantly increased basophil count ($p = 5.93 \times 10^{-$
315 ¹¹, Wilcoxon Rank Sum test), with a 4.5-fold increase in the prevalence of basophilia $>0.1 \times 10^9/L$ and
316 8.6-fold increase in the prevalence of basophilia $>0.2 \times 10^9/L$, relative to participants without a *GNBI*

317 driver mutation (13.8% vs 3.2% and 5.0% vs 0.6% for basophilia >0.1 and $>0.2 \times 10^9/L$, $n =$
318 178/431,353 for *GNBI* mutant/wild-type respectively) (Figure 4B). Individuals with *IDH2*-CH had
319 significantly lower eosinophil counts ($p = 3.63 \times 10^{-10}$, Wilcoxon Rank Sum test) and a propensity to
320 eosinopenia, with 12/92 (13.0%) participants with *IDH2*-CH having absolute eosinopenia (eosinophils
321 $= 0 \times 10^9/L$) and 45/92 (48.9%) having an eosinophil count $<0.1 \times 10^9/L$, by contrast individuals without
322 *IDH2* mutations had rates of eosinopenia of 2.9%/20.8% for absolute/ <0.1 eosinopenia respectively (n
323 $= 70/431,461$ for *IDH2* mutant/wild-type respectively) (Figure 4C). Only *IDH2*-CH demonstrated a
324 significant association between eosinophil count and clone size ($r_s = -0.51$, $p = 2.67 \times 10^{-7}$); we observed
325 no such association between *GNBI*-CH and basophil count ($r_s = 0.13$, $p = 0.09$) (Supplementary Figure
326 5), though of note basophils are the rarest of the white blood cell subsets and as a result their counts are
327 zero-biased, which may have confounded any putative association.

328

329 Discussion

330 We developed the CHIC framework and assessed a RF classifier that predicts the presence of high-risk
331 CH from just five CBC variables and an individual's age. This approach, named Clonal Haematopoiesis
332 Inference from Counts (CHIC), can discriminate between individuals with and without mutations in
333 five CH genes associated with high-risk of developing MN. Notably, CHIC retained an ability to
334 discriminate high-risk CH cases from controls even amongst individuals without cytopenias,
335 erythrocytosis or thrombocytosis, suggesting it may highlight individuals that may not otherwise come
336 to medical attention. CHIC is an important first step towards developing a scalable screening test to
337 identify individuals likely to harbour high-risk CH, who would then be prioritised for targeted NGS.
338 This would not only vastly reduce the number needed to screen (NNS) per case of high-risk CH
339 identified, but it would also justify the need to perform genetic testing. Even with its current limitations,
340 the use of CHIC with a stringent cut-off probability on individuals without cytopenia or thrombo-
341 /erythrocytosis would still markedly reduce the NNS from 727 to 40 individuals per case of high-risk
342 CH (based on the prevalence of high-risk CH in an unselected population vs in those predicted as having
343 high-risk CH by CHIC). The implementation of a scalable screening test would represent a significant
344 milestone in myeloid cancer prevention, by addressing a key bottleneck in recruitment to interventional
345 studies.

346

347 However, despite its promising metrics as a screening test, the performance of CHIC in an unselected
348 population was limited by the rarity of high-risk CH, necessitating ceding sensitivity to achieve an
349 acceptable PPV. Performance was further reduced for the restricted analysis of individuals without
350 cytopenias or thrombo-/erythrocytosis. By re-training our model in this population, we found that CHIC
351 was still able to discriminate individuals with and without high-risk CH, but the resultant small
352 reduction in AUC (0.80 vs 0.85) exacerbated the difficulty in balancing sensitivity and PPV, precluding
353 its use at population scale.

354

355 One approach for enhancing the performance of CHIC is to target its use on a population with a higher
356 prevalence of high-risk CH. CHIC was trained within the age constraints of the UKB, but since the
357 prevalence of high-risk mutations in splicing factors (*SF3B1*, *SRSF2*, *U2AF1*) rises sharply over the age
358 of 70 years, we anticipate that application of CHIC in an older population would result in improved
359 performance. Similarly, targeting CHIC to individuals with a polygenic^{15,23} or monogenic²⁴
360 predisposition to CH is also likely to improve its performance/PPV.

361

362 An alternative approach would be to integrate higher-resolution CBC data into the CHIC classifier, to
363 improve its ability to identify high-risk CH. Some of the most discriminative CBC indices for high-risk
364 CH are derived summary statistics e.g. RDW, PDW and MCH calculated from single-cell measurements
365 (i.e. RDW is a measure of variation in red cell volumes). The integration of the raw or otherwise
366 summarised single-cell measurements has the potential to improve the prediction of high-risk CH, for
367 example by revealing a fraction of cells with distinct indices arising from the CH clone or identifying
368 other characteristic patterns of variation in these measurements; such raw (or “non-classical”) CBC
369 traits have recently been exploited to explore genetic associations with blood cell morphology²⁵.

370

371 Beyond MN prevention, CH is of wider public health relevance due to its association with non-
372 haematological disorders, most notably atherosclerotic heart disease. Since *JAK2*-CH exhibits the
373 strongest association with cardiovascular outcomes²⁶, and was also the most amenable to prediction in
374 our study, we anticipate that CHIC may also have utility in the primary prevention of cardiovascular
375 disease by facilitating the identification of individuals with *JAK2*-CH. By retrofitting CH screening on
376 to a routine blood test, we believe our CHIC approach presents an important step towards scalable,
377 practical and inexpensive ML-based screening for high-risk CH and provides a proof-of-concept that
378 individuals with high-risk CH can be differentiated from those without, based on CBC indices.

379

380 **Data sharing**

381 All data used in this study are publicly available from the UK Biobank (<https://www.ukbiobank.ac.uk/>).
382 Researchers may apply for access to the UK Biobank data via the Access Management System
383 (<https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access>).

384

385 **Code availability**

386 Scripts used to query the UK Biobank dataset are available from:
387 https://github.com/IsabellaWithnell/Predicting_CH. Scripts used to implement the machine learning
388 framework described in the manuscript are available from: <https://github.com/billydunn/chic>.

389

390 **Declaration of Interests**

391 G.S.V. is a consultant to STRM.BIO and holds a research grant from AstraZeneca for research unrelated
392 to that presented here. S.W. is an employee of AstraZeneca. M.A.F. is an employee and stockholder of
393 AstraZeneca. The other authors declare no competing interests.

394

395

396

397 **Figure Legends**

398 **Figure 1: Final cohort derivation in the UK Biobank.** We selected only those with all CBC
399 parameters (that is, no missing indices) ($n = 469,739$). We apply further filtering to include only those
400 who additionally have WES data available, to facilitate identification of CH. Concurrently, we compute
401 Spearman correlation and select only one CBC parameter where two parameters exhibit high positive
402 or negative correlation ($|r_s| \geq 0.9$). Finally, we exclude individuals who were annotated as having a
403 prevalent diagnosis of a haematological malignancy, or those who had an incident diagnosis shortly
404 after recruitment/blood draw (using an arbitrary threshold of within 30 days of recruitment). This gives
405 us a final dataset of 431,531 participants, each labelled as “CH” or “no CH” for input into our
406 downstream supervised ML pipeline.

407

408 **Figure 2: Performance of machine learning classifiers to predict the presence of clonal**
409 **haematopoiesis.** Panel A shows the receiver operating characteristic (ROC) curve for classifiers
410 predicting any-driver CH using age, sex and 18 CBC parameters as features. DT = Decision Tree, RF
411 = Random Forest and XGB = eXtreme Gradient Boosting models. In each case, the ROC curve for the
412 model approximating the median AUC (area under the ROC curve) from ten repeats of model training
413 is shown. Panel B shows the performance of driver gene-specific models across all three model types
414 (DT, RF, XGB), using the same features. Boxplots are derived from the ten repeats of model building;
415 whiskers show the range of AUC values. Panel C shows the performance of RF classifiers of driver
416 gene-specific CH, using either: age and sex as the only features; or CBC features only (with age and
417 sex matching of cases to controls, to capture the predictive performance of CBC indices alone); or age,
418 sex and CBC features (without age and sex matching, thereby capturing the predictive performance of
419 CBC indices in combination with basic demographics).

420

421 **Figure 3: Optimisation, Variable Importance and Performance of a classifier of high-risk CH.**
422 Panel A shows the ROC curve for this RF model, which has been constructed and AUC calculated
423 based on performance in the unseen test set. Red, performance of model approximating the median
424 AUC. Upper and lower bounds represent performance of the models with the maximum and minimum
425 AUC from ten repeats of model training respectively. Panel B shows the impact of iterative feature
426 selection on model performance (by AUC), demonstrating that performance is stable with only six
427 features. Panel C shows variable importance (by Gini Index, scaled to the most important variable) of
428 features in our six-feature classifier. Panel D shows the trade-off between sensitivity (blue) and positive
429 predictive value (red) for this six-feature classifier.

430

431 **Figure 4: Machine learning models of driver gene CH for biological inference.** Panel A shows an
432 Upset plot generated by computing variable importance and summarising the overlap (vertical bars)
433 between the top two most important variables in RF classifiers of driver gene CH. This captures

434 expected associations (*JAK2* and *CALR* share platelet crit and platelet count as their top two variables),
435 but also unveils unexpected associations, such as the importance of basophil count for predicting the
436 presence of *GNBI-CH* and the importance of eosinophil count in predicting the presence of *IDH2-CH*.
437 PCT = platelet crit, NE = neutrophil count, PLT = platelet count, RDW = red cell distribution width,
438 EO = eosinophil count, BAS = basophil count, MCV = mean cell volume, RET = reticulocyte count,
439 LY = lymphocyte count. Panel B shows a histogram of basophil counts in carriers of *GNBI-CH* (n =
440 178) versus those without (n = 431,353); the basophil count is shifted to the right in those with *GNBI-*
441 *CH*, who have a relatively high prevalence of basophilia. Panel C shows the histogram of eosinophil
442 counts in individuals with (n = 70) and without (n = 431,461) *IDH2-CH*; there is a higher proportion of
443 absolute eosinopenia (i.e. eosinophil count = 0) in individuals with *IDH2-CH*. In both panels B and C,
444 the y axis is density, to facilitate direct comparison between imbalanced classes.

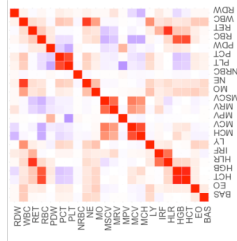
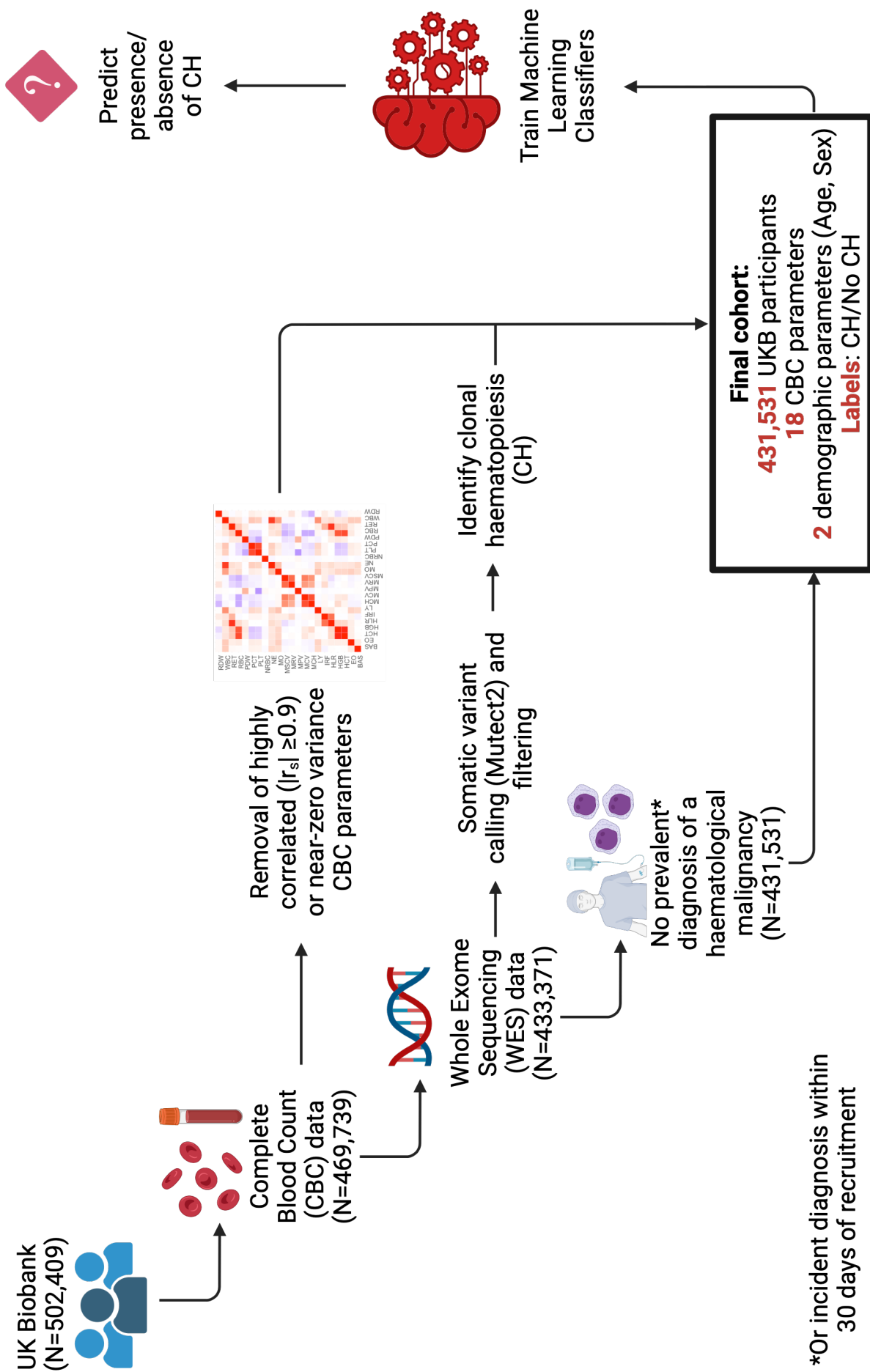
445

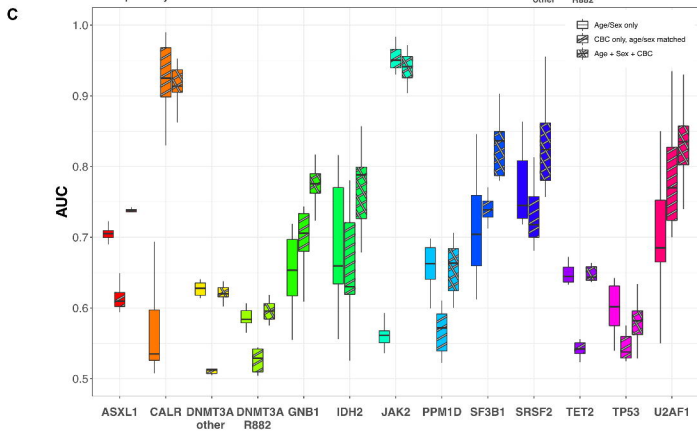
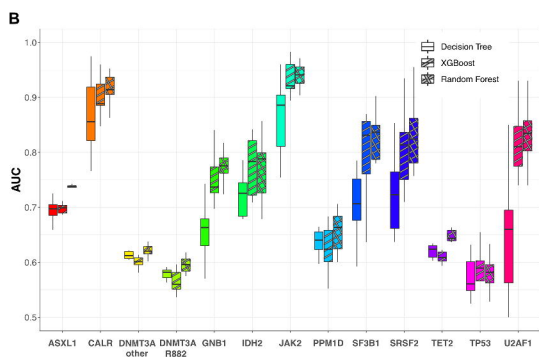
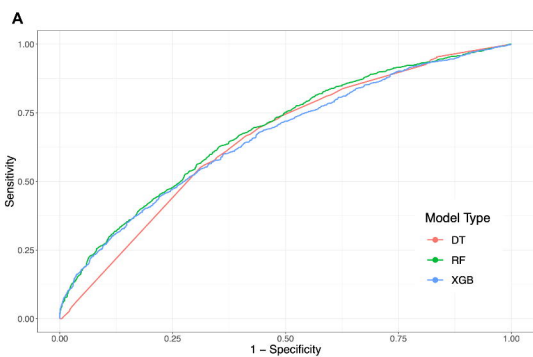
446

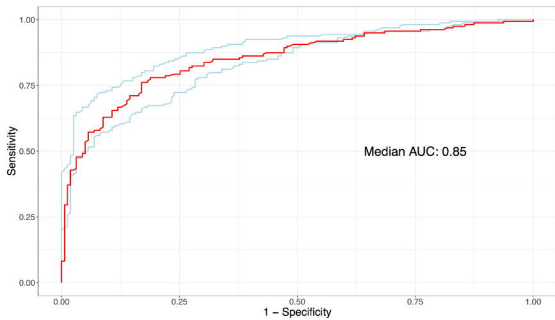
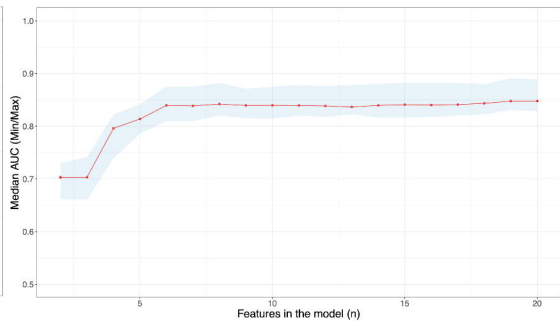
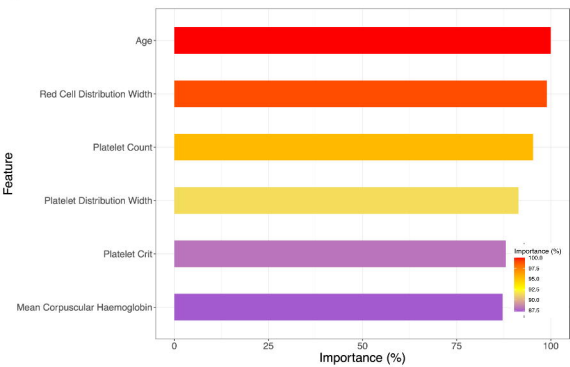
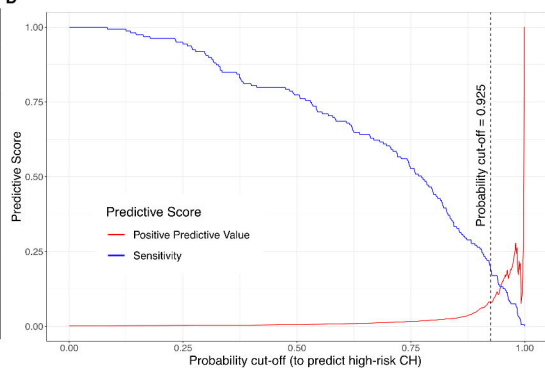
447 **References**

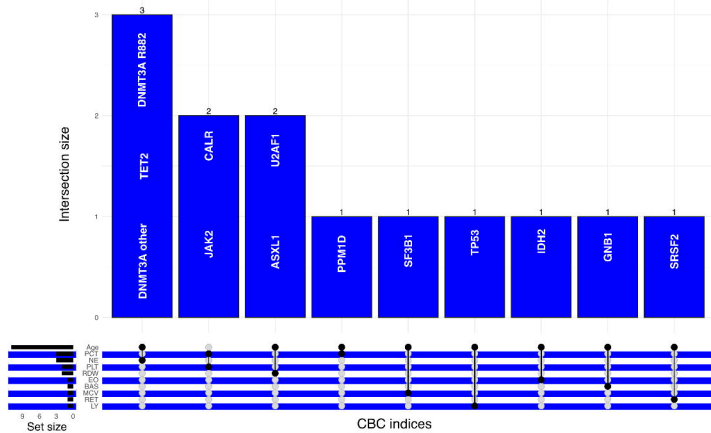
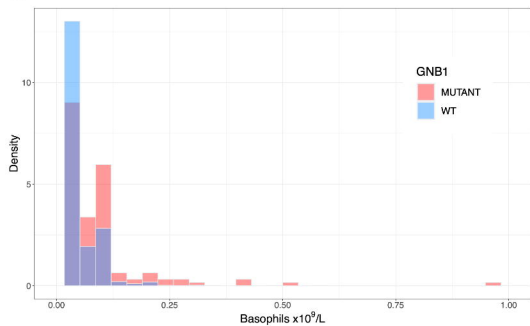
- 448 1. Harker, L. A. & Finch, C. A. Thrombokinetis in man. *J Clin Invest* **48**, 963–974 (1969).
- 449 2. Kaushansky Kenneth. Lineage-Specific Hematopoietic Growth Factors. *New England Journal of*
450 *Medicine* **354**, 2034–2045 (2006).
- 451 3. Cosgrove, J., Hustin, L. S. P., de Boer, R. J. & Perié, L. Hematopoiesis in numbers. *Trends*
452 *Immunol* **42**, 1100–1112 (2021).
- 453 4. Sender, R. & Milo, R. The distribution of cellular turnover in the human body. *Nat Med* **27**, 45–
454 48 (2021).
- 455 5. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**,
456 343–350 (2022).
- 457 6. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in
458 normal human skin. *Science* **348**, 880–886 (2015).
- 459 7. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science*
460 **362**, 911–917 (2018).
- 461 8. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations.
462 *Nature* **561**, 473–478 (2018).
- 463 9. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J*
464 *Med* **371**, 2488–2498 (2014).
- 465 10. Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA
466 sequence. *N Engl J Med* **371**, 2477–2487 (2014).
- 467 11. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and
468 malignancies. *Nat Med* **20**, 1472–1478 (2014).
- 469 12. McKerrell, T. *et al.* Leukemia-associated somatic mutations drive distinct patterns of age-related
470 clonal hemopoiesis. *Cell Rep* **10**, 1239–1245 (2015).
- 471 13. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring
472 AML-associated mutations is ubiquitous in healthy adults. *Nat Commun* **7**, 12484 (2016).
- 473 14. Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in
474 the elderly. *Blood* **130**, 742–752 (2017).

- 475 15. Kar, S. P. *et al.* Genome-wide analyses of 200,453 individuals yield new insights into the causes
476 and consequences of clonal hematopoiesis. *Nat Genet* **54**, 1155–1166 (2022).
- 477 16. Gu, M. *et al.* Multiparameter prediction of myeloid neoplasia risk. *Nat Genet* **55**, 1523–1530
478 (2023).
- 479 17. Weeks, L. D. *et al.* Prediction of Risk for Myeloid Malignancy in Clonal Hematopoiesis. *NEJM*
480 *Evidence* **2**, (2023).
- 481 18. Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**,
482 400–404 (2018).
- 483 19. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range
484 of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
- 485 20. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical*
486 *Software* **28**, 1–26 (2008).
- 487 21. Liaw, A. & Wiener, M. Classification and Regression by RandomForest. *Forest* **23**, (2001).
- 488 22. Khoury, J. D. *et al.* The 5th edition of the World Health Organization Classification of
489 Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. *Leukemia* **36**, 1703–
490 1719 (2022).
- 491 23. Kessler, M. D. *et al.* Common and rare variant associations with clonal haematopoiesis
492 phenotypes. *Nature* **612**, 301–309 (2022).
- 493 24. DeBoy, E. A. *et al.* Familial Clonal Hematopoiesis in a Long Telomere Syndrome. *N Engl J Med*
494 **388**, 2422–2433 (2023).
- 495 25. Akbari, P. *et al.* A genome-wide association study of blood cell morphology identifies cellular
496 proteins implicated in disease aetiology. *Nat Commun* **14**, 5023 (2023).
- 497 26. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N*
498 *Engl J Med* **377**, 111–121 (2017).
- 499
500





A**B****C****D**

A**B****C**