SOFTWARE TOOL ARTICLE

cophesim: a comprehensive phenotype simulator for testing novel association methods [version 1; referees: 2 approved]

Ilya Y. Zhbannikov, Konstantin G. Arbeev, Anatoliy I. Yashin

Biodemography of Aging Research Unit (BARU) at Social Sciences Research Institute (SSRI), Duke University, Durham, NC, 27705, USA

v1 **First published:** 01 Aug 2017, **6**:1294 (doi: 10.12688/f1000research.11968.1)
Latest published: 01 Aug 2017, **6**:1294 (doi: 10.12688/f1000research.11968.1)

Abstract

Simulation is important in evaluating novel methods when input data is not easily obtainable or specific assumptions are needed. We present *cophesim*, a software to add the phenotype to generated genotype data prepared with a genetic simulator. The output of *cophesim* can be used as a direct input for different genome wide association study tools. *cophesim* is available from https://bitbucket.org/izhbannikov/cophesim.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees	
	1	**2**
version 1 published 01 Aug 2017	✓ report	✓ report

1 **Arnold B. Mitnitski**, Dalhousie University, Canada

2 **Lars Rönnegård**, Dalarna University, Sweden
Elena - Flavia Mouresan, Swedish University of Agricultural Sciences, Sweden

Discuss this article

Comments (0)

Corresponding authors: Ilya Y. Zhbannikov (ilya.zhbannikov@duke.edu), Konstantin G. Arbeev (konstantin.arbeev@duke.edu)

Author roles: Zhbannikov IY: Conceptualization, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Arbeev KG**: Resources, Supervision, Writing – Review & Editing; **Yashin AI**: Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Zhbannikov IY, Arbeev KG and Yashin AI. **cophesim: a comprehensive phenotype simulator for testing novel association methods [version 1; referees: 2 approved]** *F1000Research* 2017, **6**:1294 (doi: 10.12688/f1000research.11968.1)

Grant information: This work was supported by the National Institute on Aging of the National Institutes of Health (NIA/NIH) under award numbers P01AG043352, R01AG046860, and P30AG034424
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 01 Aug 2017, **6**:1294 (doi: 10.12688/f1000research.11968.1)

Introduction

Genome-wide association studies (GWAS) are routine in population research. New methods are being developed for better accessing complex associations between genotypes and phenotypes, uncovering genotype structures or testing evolutionary hypotheses. Testing the novel methods requires experimental data, which may not be easily obtainable. One solution is to use artificial data simulated with specific assumptions.

The best existing phenotype simulators, such as: *GENOME*[1], *Plink*[2], *phenosim*[3], *CoaSim*[4], *Fregene*[5], *ForSim*[6], *QuantiNemo*[7], *GCTA*[8], *HapGen*[9], *SeqSimla*[10], and *SimRare*[11] offer qualitative and dichotomous simulated phenotype. But the known phenotype simulation software tools have some limitations, which may prevent customers from using them: (i) the majority, if not all, of the phenotype simulation software tools do not offer simulation of survival traits/time-to-event outcome, making it impossible to test respective hypotheses of associations; (ii) some of the tools are not easy to use, due to wide range of parameters, which the user has to provide and control (rather than calculate them automatically), making them unnecessarily difficult to use and preventing the user from future use of the tool; (iii) phenotype simulation is often offered as an auxiliary part of the genetic simulation routine, and therefore the user first has to perform a time-consuming unavoidable genetic simulation in order to obtain the phenotype; (iv) in situations when the genetic data is already simulated from other tools, only *phenosim* and *GCTA* offer adding simulated phenotype to such data. Consequently, it is necessary to have a new, simple and flexible phenotype simulation tool with plain algorithmic assumptions.

Consequently, we present *cophesim*, a comprehensive phenotype simulation tool that was developed to add a phenotype to corresponding genotypes simulated by other simulation tool (Table S1). *cophesim* offers simulation of continuous, dichotomous and survival traits, with different (user-provided) effect sizes of causal variants, with the ability to simulate epistatic interactions. It also can simulate phenotype within gene-environment interaction assumptions using up to 10 covariates.

Methods

Implementation

The workflow (see Figure 1) includes the following stages: (i) Input data pre-processing; (ii) phenotype simulation; (iii) generation of final output files.

Input data

Currently *cophesim* accepts the genotype output data from *Plink*, *MS*[12] and *GENOME* software applications. Phenotypes (dichotomous, continuous and survival) are then added according to the following simulation scenarios.

Dichotomous phenotype

Dichotomous phenotype for i^{th} individual ($i = 1...N$, where N is the total number of individuals in a dataset) is simulated according

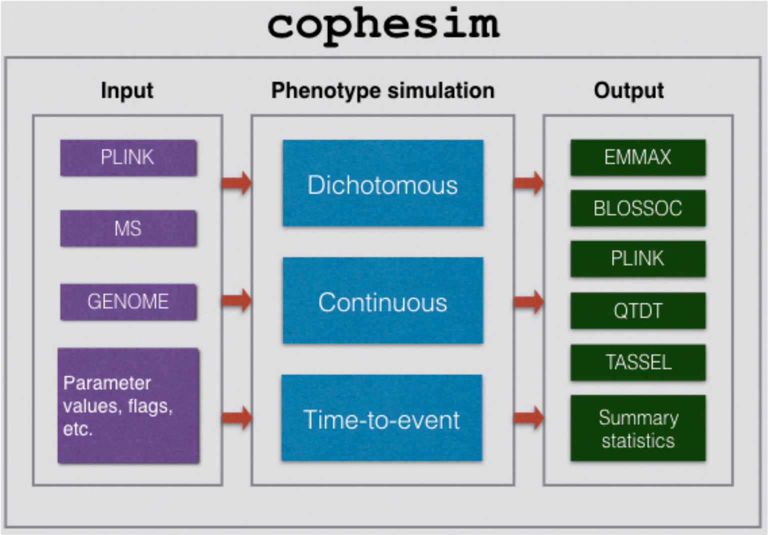

Figure 1. Workflow of *cophesim* has three stages: (1) Input stage, where the input data (can be provided in one of the three formats: *Plink*, *MS* and *GENOME*, see the user manual - Supplementary File 1) along with the other input parameters (such as causal variants with size effects, output format, etc.) is prepared for phenotype simulation; (2) Phenotype simulation stage, where different types of phenotypic traits are simulated: dichotomous, continuous and time-to-event ('survival'); (3) Output stage – the final stage, where simulated phenotype data are packed to various formats in order to be directly usable by six GWAS tools: *EMMAX*, *BLOSSOC*, *Plink*, *QTDT*, *TASSEL* and *GenABEL*. Summary statistics are generated at the output stage as well.

to the logistic model (if the user provided effect sizes for causal variants):

$$p_i = \frac{1}{1+e^{-z_i}} \qquad (1)$$

where p_i is the probability of a particular outcome. In *cophesim*, it is a probability of a "case" (cases are m~arked by "1", and "0" are controls in simulated dichotomous phenotype) for i^{th} individual. If p_i is greater than the some threshold p_0 (we use p_0 ~ $U(0, 1)$), then the phenotype for i^{th} individual is set to "1" and to "0" otherwise. The variable **z** is determined with the following equation:

$$z_i = \sum_{j=1}^{M} E_j g_{ij} + \sum_{j=1}^{K} \alpha_j X_{ij} + \epsilon_i \qquad (2)$$

E_j – effect size for j^{th} variant, user-defined; g_{ij} – value of j^{th} genetic marker for i^{th} individual; α_j - effect size for j^{th} covariate and X_{ij} is a value of j^{th} covariate for a i^{th} individual (the term $\sum_{j=1}^{K} \alpha_j X_{ij}$ is added to represent gene-environment iterations); ϵ_i – a standard normal residual, $\epsilon_i\, N(0, 1)$, computed for i^{th} individual, M is a total number of genetic variants and K is a total number of covariates used.

If the user did not provide the effect sizes for causal variants, the following strategy is then used:

$$z_i = \sum_{j=1}^{M} w_{ij} + \sum_{j=1}^{K} \alpha_j X_{ij} + \epsilon_i \qquad (3)$$

Here w_{ij} is a weight and computed as follows: $w_{ij} = \frac{g_{ij} - 2MAF_j}{(2MAF_j(1-MAF_j))^{1/2}}$ (a standardization procedure, and the matrix W containing element w_{ij} is called a standardized genotype matrix[8]; MAF_j – a minor allele frequency for j^{th} genetic variant, and the other values are the same as described above. This strategy allows using defined genetic architecture in a simulated population.

Continuous phenotype

Qualitative (continuous) phenotype for i^{th} individual is simulated according to the linear regression scenario according to the equations (2) or (3) (in case if effect sizes were not supplied).

Inverse Probability method

We model a survival phenotype from the proportional hazards model using the inverse probability method[13]: if U is uniform in (0, 1) and if $S(\cdot|\mathbf{z})$ is the conditional survival function derived from the proportional hazards model: $S(t|\mathbf{z}) = e^{-H_0(t)e^{\mathbf{z}}}$, then the random variable

$$T = S^{-1}(\cdot\,|\,\mathbf{z}) = H_0(t)^{-1}\left(\frac{-log(U)}{e^{\mathrm{z}}}\right) \qquad (4)$$

has survival function $S(\cdot|\mathbf{z})$. In this equation, $H_0(t)$ is a cumulative baseline hazard. By default, we use the Weibull cumulative baseline hazard: $H_0(t) = \lambda\rho t^{\rho-1}$; **z** is the same parameter that defined above, for each individual, and depends on whether the user provided effect sizes for causal variants or not. We also implemented exponential and Gompertz hazards.

Linkage Disequilibrium

The simplest way to simulate collinearity between two SNPs, g_1 and g_2, with effect sizes E_1 and E_2 is to replace some portion of g_2 with g_1 values according to provided r_{12}^2 coefficient, which reflects a correlation between two SNPs. We also consider applying other techniques, such as copulas, in order to simulate LD.

Epistatic interactions

These are modeled with the following equation for i^{th} individual:

$$z_i = E_1 g_{1i} + E_2 g_{2i} + E_{12} g_{1i} g_{2i} + \sum_{j=1}^{k} \alpha_i X_{ji} + \epsilon_i \qquad (5)$$

where the term $E_{12}\, g_{1i}\, g_{2i}$ is the interaction term in which E_{12} is the epistatic effect size (user-defined, zero by default); α_j is the effect size for j^{th} covariate X.

Output files

Output files are in the formats as the direct inputs for the following tools: *EMMAX*[14], *Blossoc*[4], *Plink* (.ped file), *QTDT*[15], *TASSEL*[16], *GenABEL*[17] (see Table 1).

Operation

cophesim is freely available for download from the following link: https://bitbucket.org/izhbannikov/cophesim. Requirements: *Python* v2.7.10 and newer, *plinkio* v0.9.6, R v3.2.4 and newer, *Plink* v1.07, - in order to run the examples. The user manual is provided in a separate file "cophesim.pdf" located in the program directory and is also available as Supplementary File 1.

Table 1. Output file formats supported by phenotype simulator *cophesim*. Applying one of the options shown below controls the output format. Each output format has a special suffix type, which defines the file format. These output formats are concordant to those used in *phenosim*.

Application	Option	Commentary
EMMAX	-emmax	Suffices .emma_geno, .emma_pheno
BIOSSOC	-blossoc	Suffices .blossoc_pos, .blossoc_geno
PLINK	-plink	Used by default across all phenotypes, except survival. Suffices .ped, .map, .pheno.
QTDT	-qtdt	Suffices .ped, .map, .dat
TASSEL	-tassel	Suffices .poly, .trait
GenABEL	-	This format is used in simulation of survival phenotype.

Use case

Below we present an example that shows simulation of genetic data and then simulation of three different phenotypic traits. Other examples and installation instructions are provided at the program website and also in the user manual. Refer to the user manual for description of input parameters.

```
#----------------Example begins------------------#
#Step 1: genetic data simulation:
plink --simulate-ncases 5000 --simulate-ncontrols
   5000 --simulate wgas.sim --out sim.plink --make-
bed
#Step 2: Convert .bed to .ped:
plink --bfile sim.plink --recode --out sim.plink
#Step3: phenotype simulation from previously made
   genetic data:
python cophesim.py -i sim.plink -o testout -itype
   plink -otype plink -c -ce effects.txt -s -gomp
#-----------------End of example-----------------#
```

In this example, we first (Step 1) simulate genetic data using *Plink*. We simulate *N.cases* = *N.control* = 5,000 cases and controls and 1,000 SNPs (defined in `wgas.sim` file, refer to the *Plink* website to see documentation for this type of file). Then (Step 2) we convert a binary `sim.plink.bed` file to `sim.plink.ped` (option `--recode` in Plink). This step is not required since cophesim can handle binary *Plink* files (`.bed`, `.bim`, `.fam`), but we provide this step in order to show the ability of the program to deal with *Plink* PED format. Finally (Step 3), we simulate dichotomous (by default), continuous (option `-c`) and survival (option `-s`) traits from previously simulated data stored in files `sim.plink.ped` and `sim.plink.map`. Note that we simulate survival trait with Gompertz hazard function (option `-gomp`); effect sizes for causal variants are provided in file `effects.txt` (to include this file we use option `-ce`).

ROC curves

We provide Receiver-Operating Characteristic (ROC) curves (Figure 2) constructed from association tests performed on a simulated dataset. Simulation and association testing were performed with *Plink* suite. The following parameters were used: *N* = 10,000 individuals, *N.snp.c* = 100 causal, with total *N.snp* = 1,000 variants. Causal variants were labeled with ‘1’ and the other (neutral) variants were labeled with ‘0’. These labels are then used later as true identifiers during calculation of TPR (true positive rate) and FPR (false positive rate). Dichotomous, continuous and survival phenotypic traits were simulated with *cophesim*. Then association tests were performed with *Plink* for dichotomous and continuous traits (using *Plink* flags `-logistic` and `-regression`, respectively). Association tests for survival trait were performed with the R package *GenABEL*. Then calculated *p*-values provided by association tests for each variant were compared to the significance threshold. Those variants passed the threshold were recognized as causal and associated with simulated

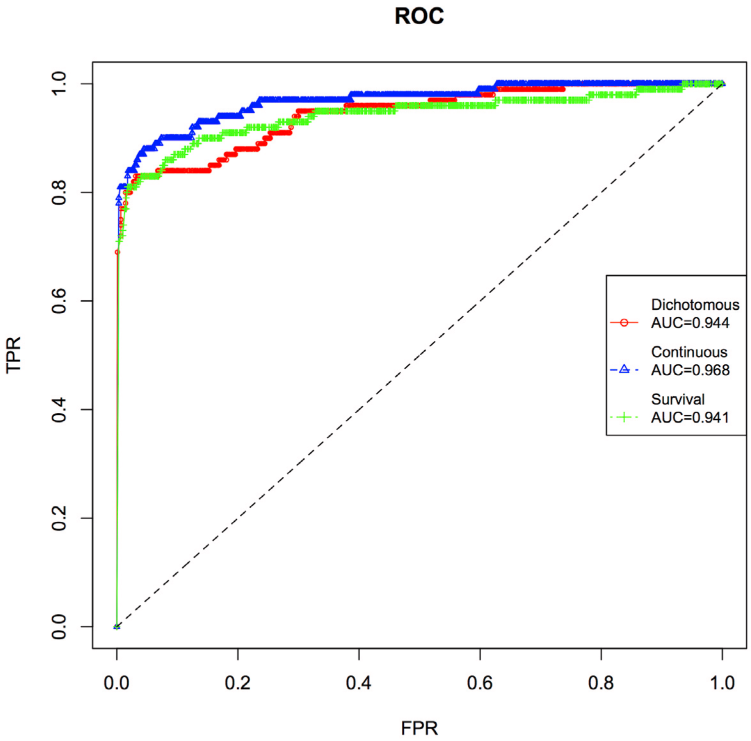

Figure 2. ROC curves constructed from results of association tests performed on a simulated dataset of *N* = 10,000 individuals, 100 causal and 1,000 of total SNP sites. TPR: True Positive Ratio, FPR: False Positive Ratio. These results were calculated for dichotomous, continuous and survival traits. The dashed, 45 degrees line represents random guessing.

phenotype. These classification results later were compared to the true identifiers (defined above) in order to obtain TPR and FPR. For all these tests, we varied the significance threshold from 0 to 1 with the increment of 0.001.

The R code to construct ROC curves is provided in the file "roc.R". This file is attached to this computer note and also in the data repository: https://bitbucket.org/izhbannikov/cophesim_data/ROC/roc.R

Conclusion

In this work we presented the *cophesim* for phenotype simulation from genetic data obtained either from simulation or real data collecting. *cophesim* makes it possible to simulate various demographic models under user-defined scenarios.

Software and data availability

Tool and source code available from: https://bitbucket.org/izhbannikov/cophesim

Archived source code as at time of publication: doi:10.5281/zenodo.810195[18]

License: MIT

The example script and output files for the software are available at: https://doi.org/10.5281/zenodo.804090[19].

To test the *cophesim* we provided a repository "cophesim_data": https://bitbucket.org/izhbannikov/cophesim_data. Download or clone this repository to be able to run tests.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by the National Institute on Aging of the National Institutes of Health (NIA/NIH) under award numbers P01AG043352, R01AG046860, and P30AG034424.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIA/NIH.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

Table S1: Best available phenotype/genotype simulation software applications and their comparison to *cophesim* in terms of ability to simulate different types of phenotypic traits. (https://f1000researchdata.s3.amazonaws.com/supplementary/11968/c65c7dddd305-4043-a722-e850f2413f10.docx)

Supplementary File 1: User manual for *cophesim* (https://f1000researchdata.s3.amazonaws.com/supplementary/11968/42ab5de2-8130-4b8c-a7ce-abb2f3d55648.pdf).

References

1. Liang L, Zöllner S, Abecasis GR: **Genome: a rapid coalescent-based whole genome simulator.** *Bioinformatics.* 2007; **23**(12): 1565–7.
PubMed Abstract | **Publisher Full Text**

2. Purcell S, Neale B, Todd-Brown K, *et al.*: **Plink: A tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet.* 2007; **81**(3): 559–575.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

3. Günther T, Gawenda I, Schmid KJ: **phenosim--A software to simulate phenotypes for testing in genome-wide association studies.** *BMC Bioinformatics.* 2011; **12**(1): 265.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

4. Mailund T, Schierup MH, Pedersen CN, *et al.*: **Coasim: A flexible environment for simulating genetic data under coalescent models.** *BMC Bioinformatics.* 2005; **6**(1): 252.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

5. Hoggart CJ, Chadeau-Hyam M, Clark TG, *et al.*: **Sequence-level population simulations over large genomic regions.** *Genetics.* 2007; **177**(3): 1725–1731.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

6. Lambert BW, Terwilliger JD, Weiss KM: ***Forsim*: a tool for exploring the genetic architecture of complex traits with controlled truth.** *Bioinformatics.* 2008; **24**(16): 1821–2.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

7. Neuenschwander S, Hospital F, Guillaume F, *et al.*: **quantinemo: an individual-based program to simulate quantitative traits with explicit genetic architecture in a dynamic metapopulation.** *Bioinformatics.* 2008; **24**(13): 1552–3.
PubMed Abstract | **Publisher Full Text**

8. Yang J, Lee SH, Goddard ME, *et al.*: **Gcta: A tool for genome-wide complex trait analysis.** *Am J Hum Genet.* 2011; **88**(1): 76–82.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

9. Spencer CC, Su Z, Donnelly P, *et al.*: **Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip.** *PLoS Genet.* 2009; **5**(5): e1000477.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

10. Chung RH, Shih CC: **SeqSIMLA: a sequence and phenotype simulation tool for complex disease studies.** *BMC Bioinformatics.* 2013; **14**(1): 199.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

11. Li B, Wang G, Leal SM: **Simrare: a program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits.** *Bioinformatics.* 2012; **28**(20): 2703–4.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

12. Ewing G, Hermisson J: ***MSMS*: a coalescent simulation program including recombination, demographic structure and selection at a single locus.** *Bioinformatics.* 2010; **26**(16): 2064–5.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Bender R, Augustin T, Blettner M: **Generating survival times to simulate Cox proportional hazards models.** *Stat Med.* 2005; **24**(11): 1713–1723.
PubMed Abstract | Publisher Full Text

14. Kang HM, Sul JH, Service SK, *et al.*: **Variance component model to account for sample structure in genome-wide association studies.** *Nat Genet.* 2010; **42**(4): 348–54.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Abecasis GR, Cardon LR, Cookson WO: **A general test of association for quantitative traits in nuclear families.** *Am J Hum Genet.* 2000; **66**(1): 279–292.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Bradbury PJ, Zhang Z, Kroon DE, *et al.*: **TASSEL: software for association mapping of complex traits in diverse samples.** *Bioinformatics.* 2007; **23**(19): 2633–5.
PubMed Abstract | Publisher Full Text

17. Aulchenko YS, Ripke S, Isaacs A, *et al.*: **GenABEL: an R library for genome-wide association analysis.** *Bioinformatics.* 2007; **23**(10): 1294–6.
PubMed Abstract | Publisher Full Text

18. Zhbannikov I: **izhbannikov/release-1.4.1.** *Zenodo.* 2017.
Data Source

19. Zhbannikov I: **izhbannikov/cophesim_data: First release.** *Zenodo.* 2017.
Data Source

Open Peer Review

Current Referee Status: ✓ ✓

Version 1

Referee Report 18 September 2017

doi:10.5256/f1000research.12938.r25816

✓

Lars Rönnegård [1], **Elena - Flavia Mouresan** [2]
[1] Dalarna University, Falun, Sweden
[2] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

Zhbannikov and co-workers present the *cophesim* software that they have developed for simulating phenotypic data using genotype information. The input and output file formats are compatible with many of the most commonly used computer programs for genome wide association studies. The software is flexible, well documented and fills a gap in existing tools, especially for simulating time-to-event phenotypes. The paper is well written and easy to follow, and we only have some minor comments and suggestions.

Minor comments

- Closing bracket missing in the sentence below equation (3)
- In equation (4), if the user does not provide gene effects then the phenotype is built by the sum of the standardized genotypes for each individual. Could you motivate this choice a bit and explain why it would be useful?
- In equation (5), the subscripts look wrong. a_i should be a_j
- In the Linkage Disequilibrium section the term “copula” is used. We do not think most readers of this paper can be expected to be acquainted with copulas and a reference is needed.

Consider adding a short paragraph where you discuss limitations and the possibility to add further functionality in the future, including:

- Dominance effects
- Probit link for binary data
- Simulation of correlated traits
- Alternative ways to simulate LD including a copula approach

Check that the following link (at the end of the paper) works: https://bitbucket.org/izhbannikov/cophesim_data/ROC/roc.R (we were able to retrieve the code from https://bitbucket.org/izhbannikov/cophesim_data/src/)

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 08 August 2017

doi:10.5256/f1000research.12938.r24718

✓ **Arnold B. Mitnitski**
Department of Medicine, Dalhousie University, Halifax, NS, Canada

In the manuscript “Cophesim: a comprehensive phenotype simulator for testing novel association methods”, I. Zhbannikov and colleagues from the Duke University, presented a software that allowed to generate genotype data prepared with a genetic simulator for the use in the investigations of the genome wide association study (GWAS) tools. The rational for development of the software is clearly explained. The idea of the study is to use computer simulations to model data with specific assumptions. Similar simulators are known but all of them do not allow simulate survival. There are several other disadvantage with the existing simulators reviewed by the authors.

The description of the software is technically sound. The methods section is clearly presented. Dichotomous phenotype are simulated according to the logistic model with the covariates being genetic variants and covariates. Continuous phenotypes are simulated using the linear regression. Survival phenotype is modeled using the proportional hazards with inverse probability method.

The details of the code, methods and analysis allow replication of the software and its use by the others. The methods section is clearly presented. Dichotomous phenotype are simulated according to the logistic model with the covariates being genetic variants and covariates. The output formats are compatible with the other applications (Table 1). It is useful example if using the simulator and the other examples are available in the manual. The ROC curve example is also very useful. The information provided is quite sufficient to allow interpretation of the expected results.

In short, the Cophesim is a useful tool that can be helpful in the genetic analyses. The article is scientifically sound, the methods are described with details – this article will greatly help the researcher interested in the application genetic analyses.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.