16ª ed

MIM

Penetrance estimation and clinical, demographic and genetic characterization of HDGC families sharing a founder variant followed at Centro Hospitalar Universitario São João (CHUSJ)

João Abel Rainho Fonseca

MESTRADO EM
INFORMÁTICA MÉDICA
2º CICLO DE ESTUDOS

SET|2023

16ª ed

MIM

Penetrance estimation and clinical, demographic and genetic characterization of HDGC families sharing a founder variant followed at Centro Hospitalar Universitario São João (CHUSJ)

João Abel Rainho Fonseca

MESTRADO EM
INFORMÁTICA MÉDICA
2º CICLO DE ESTUDOS

ORIENTADORES:

Carla Isabel Gonçalves de Oliveira – i3S / FMUP
Susana Alves Seixas – i3S
Ana Rita Barbosa Matos

SET|2023

Acknowledgements

Em primeiro lugar, gostaria de dirigir um agradecimento à direção do programa de metrado em Informática Médica, por me ter permitido a realização e concretização deste projeto no âmbito do mesmo, e agradecer a oportunidade de formação.

O meu primeiro agradecimento é dirigido à Doutora Carla Oliveira, orientadora deste projeto de tese e sem a qual não teria chegado a este ponto. Agradeço por todas as horas dispensadas na incansável vontade de formar e transmitir conhecimento, por toda a simpatia e pela oportunidade de trabalhar num projeto único. Foi de facto inspirador trabalhar sob a orientação da Carla e um marco no meu caminho académico.

Agradeço também à Mestre Rita Matos por todo apoio ao longo deste ano e pela partilha de todos os momentos que trilhámos ao longo deste trabalho. Por vezes um sinuoso caminho por todos os dados que tivemos de "mastigar", mas foi um caminho de aprendizagem e que chegou a um bom porto. Foi essencial a orientação da Rita ao longo deste projeto.

Um agradecimento também à Doutora Susana Seixas pela co-orientação neste projeto, sempre disponível em dar o seu contributo e orientação na sua área de especialidade.

Um agradecimento profundo à Enfermeira Luzia Garrido, que durante todo o ano aturou um aluno de mestrado que muito lhe inquiri, diversas vezes as mesmas perguntas, mas esteve sempre disponível e com um grande sorriso para fornecer toda a ajuda necessária e valiosa durante o ano todo. Ela acompanhou de perto o processo dos doentes que fizeram parte deste estudo, e só me resta dizer que a sua ética profissional é uma inspiração. Sem a Luzia nada deste trabalho seria possível.

Agradeço também a toda a equipa do serviço de Genética Médica e Centro da Mama do Centro Hospitalar São João do Porto, sem o qual não seria possível ter realizado este trabalho.

Agradeço também a todo os membros grupo *Expression Regulation in Cancer*, i3S, por todas a disponibilidade e receção, proporcionando um ambiente de trabalho saudável, oferecendo sempre ajuda nos momentos em que precisei.

Por fim, resta-me agradecer a toda a gente que me acompanhou ao longo desta viagem. À Teresa que sempre esteve comigo nos momentos mais felizes e nos mais difíceis deste ano, contribuindo sempre naquilo que podia. Aos amigos do café, que muito ajudaram não só na descompessão, mas também na discussão valiosa da ciência no café, obrigado Alexandre, Rita e Carla. À minha família que esteve sempre presente e demonstrou interesse no meu projeto, mas acima de tudo, por demonstrarem interesse na minha felicidade e contribuírem para tal. Como sempre. Obrigado.

Abstract

Hereditary diffuse gastric cancer (HDGC), caused by CDH1 germline pathogenic, or likely-pathogenic variants, predisposes to early onset diffuse gastric (DGC) and lobular breast cancer (LBC). The *CDH1* c.1901C>T pathogenic variant was previously defined as founder variant in the Northern Portuguese population, reported among different families, sharing a common ancestor. The main goal of this thesis is to characterize these HDGC families, through establishment of extended family pedigrees and an updated comprehensive database with relevant data, to be used in disease risk calculation for carriers of this variant. Secondly, we aim to explore the transcriptomic landscape of selected cases with different DGC stages to uncover genes and pathways disrupted in HDGC setting, thus becoming targets for possible therapeutics, or markers in surveillance genetic testing.

Detailed clinical and demographic data of individuals from 9 known families carrying the c.1901C>T variant followed at Centro Hospitalar Universitário São João (CHUSJ) were compiled, followed by the development of a reference database with updated clinical records. Pedigrees were built using an open-source software, subsequently trimmed for penetrance estimation analysis. Lifetime risk estimations were performed resorting to Kaplan-Meier risk curves and the Genotype-Restricted Likelihood (GRL) method.

Extended pedigree information was obtained for the 9 families through genetic testing in several hospitals and/or patient reports at genetic counselling appointments, resulting in a clinical database comprising a total of 400 individuals, with 13.0% (52/400) cases of gastric cancer (GC) and 5.8% (12/208) of breast cancer (BC). Specifically, for the *CDH1* c.1901C>T variant, 42.8% (171/400) individuals underwent genetic testing at CHUSJ, of whom 43.8% (75/171) were found positive. Penetrance estimations by the GRL method with the Weibull model revealed 13.8% of cumulative risk by the age of 65 years for DGC, and 11.8% for LBC.

This work represents a steppingstone in the first intrafamilial lifetime-risk estimations for carriers of a single founder variant (CDH1 c.1901C>T) in the Portuguese population. It is projected that this robust clinical database will guide current and future studies on genomic and transcriptomic landscape of these HDGC families, with the goal to produce results that will contribute for guidelines' development with implications in patients' management.

Resumo

A síndrome do cancro gástrico difuso hereditário (CGDH), causado por variantes patogénicas germinativas no gene CDH1, predispõe o desenvolvimento precoce de cancro gástrico difuso (CGD) e cancro lobular da mama (CLM). A variante genética c.1901C>T deste gene foi reportada como variante patogénica fundadora numa população do norte de Portugal, partilhada por diversas famílias. O principal objetivo deste projeto é a caracterização genética e clínica destas famílias, através da criação de pedigrees das mesmas e de uma base de dados robusta e atualizada, para o desenvolvimento de cálculos de risco de doença para os portadores desta variante. Adicionalmente, pretendemos explorar o perfil transcriptómico de CGD em vários estadios para recolher gene e vias enriquecidas que estejam afetados num perfil avançado de cancro, sendo reconhecidos como potenciais alvos de terapêutica, ou biomarcadores em prevenção de CGDH.

Foram compilados dados demográficos e clínicos de indivíduos de 9 famílias com portadores conhecidos da variante c.1901C>T seguidas no Centro Hospitalar Universitário São João (CHUSJ), seguindo-se o desenvolvimento de uma base de dados de referência com registos clínicos atualizados. Foram elaborados pedigrees recorrendo-se ao uso de software open-source e adaptados para uma análise de estimativas de penetrância. Análises de estimativas de risco foram realizadas com o auxílio do método *Genotype-Restricted Likelihood (GRL)*.

Informações para os pedigrees foram obtidas através de testes genéticos em diversos hospitais e/ou de relatos de pacientes em consultas de oncogenética em diversos institutos, compilando um total de 400 indivíduos, com 130.0% (52/400) de casos de cancro gástrico e 5.8% (12/208) de casos de cancro da mama. Para a variante CDH1 c.1901C>T, 42.8% (171/400) dos indivíduos foram testados no CHUSJ, dos quais 43.8% (75/171) foram dados como positivos. As análises de penetrância pelo modelo GRL com a função de *Weibull* revelaram estimativas de risco cumulativo de 13.8% para CGD, e 11.8% para CLM à idade de 65 anos.

Este trabalho apresenta-se como uma base para o estabelecimento de estimativas de risco intrafamiliar, sendo a primeira para portadores da variante fundadora CDH1 c.1901C>T numa população portuguesa. É esperado que a base de dados resultante seja um pilar importante para estudos futuros de genómica e transcriptómica em famílias de CGDH, com o objetivo de contribuir no estabelecimento de guidelines mais conservadoras e personalizadas para a prevenção de CGDH.

Preamble

This work was accomplished within the Life-time Risk Estimations and Genetic Modifiers Of Hereditary Diffuse Gastric Cancer (LEGOH) project, funded by Fundação para a Ciência e Tecnologia (FCT) – PTDC/BTM-TEC/6706/2020.

I was awarded with a BSc research fellowship to work within this project and develop my master's thesis.

This work was presented in an oral presentation at the *Sociedade Portuguesa de Genética Humana* at Coimbra, November 2022, and as a poster presentation at the European Society of Human Genetics, at Glasgow, June 2023.

Table of Contents

Tables Index

Figures Index

Acronyms

ACMG/AMP	The American College of Medical Genetics and the Association for Molecular Pathology
ANOVA	Analysis of Variance
BC	Breast cancer
BMI	Body Mass Index
BP	Biological Process
BRCA1/2	Breast Cancer Associated gene 1/2
BRRM	Bilateral Risk Reducing Mastectomy
CC	Cellular Cmponent
CDH1	Cadherin-1
CHUSJ	Centro Hospitalar e Universitário de São João
CI	Confidence Interval
CTNNA1	Catenin alpha-1
DE	Differential Expression
DGC	Diffuse Gastric Cancer
ECIS	European Cancer Information System
ERK1/2	Extracellular signal-Regulated Kinase 1/2
F	Family
FCT	Fundação para a Ciência e a Tecnologia
FDR	False Discovery Rate
FFPE	Formalin-Fixed Paraffin-Embedded
GC	Gastric Cancer
GCO	Global Cancer Observatory
GO	Gene Ontology
GRL	Genotype Restricted Likelihood

HDGC	Hereditary Diffuse Gastric Cancer
ID	Identification
IGC	Intestinal Gastric Cancer
IGCLC	International Gastric Cancer Linkage Consortium
IHC	Imunohistochemical
IPATIMUP	Instituto de Patologia e Imunologia Molecular da Universidade do Porto
LBC	Lobular Breast Cancer
LEGOH	Life-time Risk Estimations and Genetic Modifiers Of Hereditary Diffuse Gastric Cancer
LIN	Lobular Intraepithelial Neoplasia
lncRNA	Long non-coding RNA
MAPK	Mitogen-Activated Protein Kinase
MF	Molecular Function
MHL1	MutL Homolog 1
MRI	Magnetic Resonance Imaging
MSH2/6	MutS Homolog 2/6
MUC6	Mucin-6
NGS	Next Generation Sequencing
NP	Nonparametric
OC	Obligate Carrier
PALB2	Partner and Localizer of BRCA2
PCA	Principal Component Analysis
PSCA	Prostate Stem Cell Antigen
RNA	Ribonucleic Acid
RRG	Risk Reducing Gastrectomy
RRM	Risk Reducing Mastectomy
SEER	Surveillance, Epidemiology and End Results Program
SIR	Standard Incidence Ratio
SRC	Signet Ring Cells

TAC	Transcriptome Analysis console
TNM	Tumour, Node, Metastasis Classification
TP53	Tumour Protein 53
UCA1	Urothelial Cancer Associated-1
UnkH	Unknown Histological classification

1. Introduction

1.1. Hereditary Diffuse Gastric Cancer

According to data from the Global Cancer Observatory (GCO), Gastric Cancer (GC) is the 5th most common cancer worldwide and the 4th most common cause of oncological-related mortality[1,2]. GC encompasses two main types of disorders, of which Diffuse Gastric Cancer (DGC) correlates to around 40%[3,4].

A familial setting in GC corresponds to roughly 10% of the reported cases, from which hereditary cancer syndromes represent approximately 1-to-3%[5–7]. One of the most common hereditary syndromes is Hereditary Diffuse Gastric Cancer (HDGC), an autosomal dominant syndrome characterized by the predisposition to early-onset development of not only DGC, but also of lobular breast cancer (LBC)[8,9]. Additionally, cleft lip/palate is also associated with HDGC[10,11]. Early onset of DGC is observed in approximately 10% of gastric cancers and is a phenomenon associated with HDGC, and with an aggressive manifestation before 45 years of age[7,12].

This syndrome is mostly associated with pathogenic germline variants in the *CDH1* gene[13–17], that encodes for the calcium-dependant transmembrane glycoprotein E-cadherin, essential in cell-cell adhesion in epithelial tissue [18]. In patients with HDCG, E-cadherin is lost or with a reduced expression in precursor lesions, usually due to inactivation of the second allele, leading to GC development[19–21].

E-cadherin is expressed in most epithelial tissue cells and is a known tumour suppressor protein for DGC[13,14,19]. Its interaction with α- and β-catenin (the latter being a known proto-oncogene[22–24]) is essential for cell-cell adhesion. Additionally, E-cadherin is also a strong element in establishing cell polarity and signalling, tissue architecture, as well as in mediating mechanisms such as proliferation and differentiation[18,24]. Reduced or supressed expression of this protein leads to loss of tissue integrity, resulting in poorly differentiated cells and in an increase of cell mobility and invasiveness[24–26]. The spread of these defective cells leads to *linitis plastica* event – the thickening and stiffness of the gastric wall[15].

In 1998, Gilford *et al* described a genetic cause for DGC, following the study of a Māori family with multigenerational early-onset DGC, presenting a strong linkage between *CDH1* germline variants and disease[13]. Today, *CDH1* pathogenic or likely pathogenic variants are correlated with around 40% of all HDGC reported cases, although the molecular cause of several familial cases is still unsolved [5,27,28]. It is known that some HDGC families also display germline variants in the α-catenin encoding gene, *CTNNA1*, which leads to consider this gene association to HDGC, suggesting a different HDGC gene association other than *CDH1* [29,30]. In line with this, several cancer-related genes variants have been identified in HDGC patients, such as *PALB2*, *BRCA1*, *BRCA2*, *MSH2*, but none of these demonstrate a sufficiently strong association with DGC or LBC development[28,31].

1.1.1. Diffuse Gastric Cancer (DGC)

Gastric cancer is subdivided into two main subtypes: intestinal gastric cancer (IGC) and diffuse gastric cancer (DGC)[32]. To classify GC stages, the TNM classification from the 8th edition of the American Joint Committee on Cancer is in use[33]. This classification is subdivided into tumour size (T), lymph nodes metastasis extension (N), and the existence of distant metastasis (M), which is further detailed in table 1.

Table 1. TNM classification for gastric cancer. *Adapted from* the **8th Edition AJCC Cancer Staging Manual** (AJCC, **2017**)[33]

T Category	
TX	Primary tumor cannot be assessed
T0	No evidence of primary tumor
Tis	Carcinoma in situ: intraepithelial tumor without invasion of the lamina propria
T1	Tumor invades lamina propria, muscularis mucosae, or submucosa
T1a	Tumor invades lamina propria or muscularis mucosae
T1b	Tumor invades submucosa
T2	Tumor invades muscularis propria
T3	Tumor penetrates subserosa connective tissue without invasion of visceral peritoneum or adjacent structures
T4a	Tumor invades serosa (visceral peritoneum)
T4b	Tumor invades adjacent structures and organs
N Category	
NX	Regional lymph node(s) cannot be assessed
N0	No regional lymph node metastasis
N1	Metastasis in 1 to 2 regional lymph nodes
N2	Metastasis in 3 to 6 regional lymph nodes
N3	Metastasis in 7 or more regional lymph nodes
M Category	
MX	Distant metastasis cannot be assessed
M0	No distant metastasis
M1	Distant metastasis

The gastric cancer associated with HDGC is the DGC. This type of cancer, differently from IGC, occurs most often in younger individuals[7,12,34]. Advanced DGC in an HDGC setting is associated with poor prognosis and often detected at later stages of disease. It is characterized by the diffuse infiltration of cells with poor differentiation that originate from the gastric epithelium and become invasive, leading to proliferation through the mucosa, submucosa and the muscle layers (Figure 1, panel 5), often designated as ≥pT2 tumours[32,35,36].

There are some precursor lesions associated with DGC. In a premalignant stage, signet ring cells (SRCs) can be found *in-situ* or in a pagetoid spread (pTis lesion) below a layer of epithelial cells but still restricted within the basement membrane (Figure 1, panel 2)[35–38]. In early DGC, the tumour

invades the lamina propria (pT1a), with SRCs becoming increasingly poorly differentiated (Figure 1, panels 3 and 4)[34,35]. These early lesions may be difficult to detect at this stage, as they are usually small foci of intramucosal SRCs, one of the reasons behind the usually late diagnosis of the disease[39]. In immunohistochemistry essays in tissue from HDGC patients, these early neoplastic lesions, such as Tis and T1 have been shown to have "indolent" features without expression of Ki-67 and p53, while advanced tumours exhibit pleomorphic cells with reaction to these markers[40]. These findings can be key to understand progression to advanced gastric cancer.

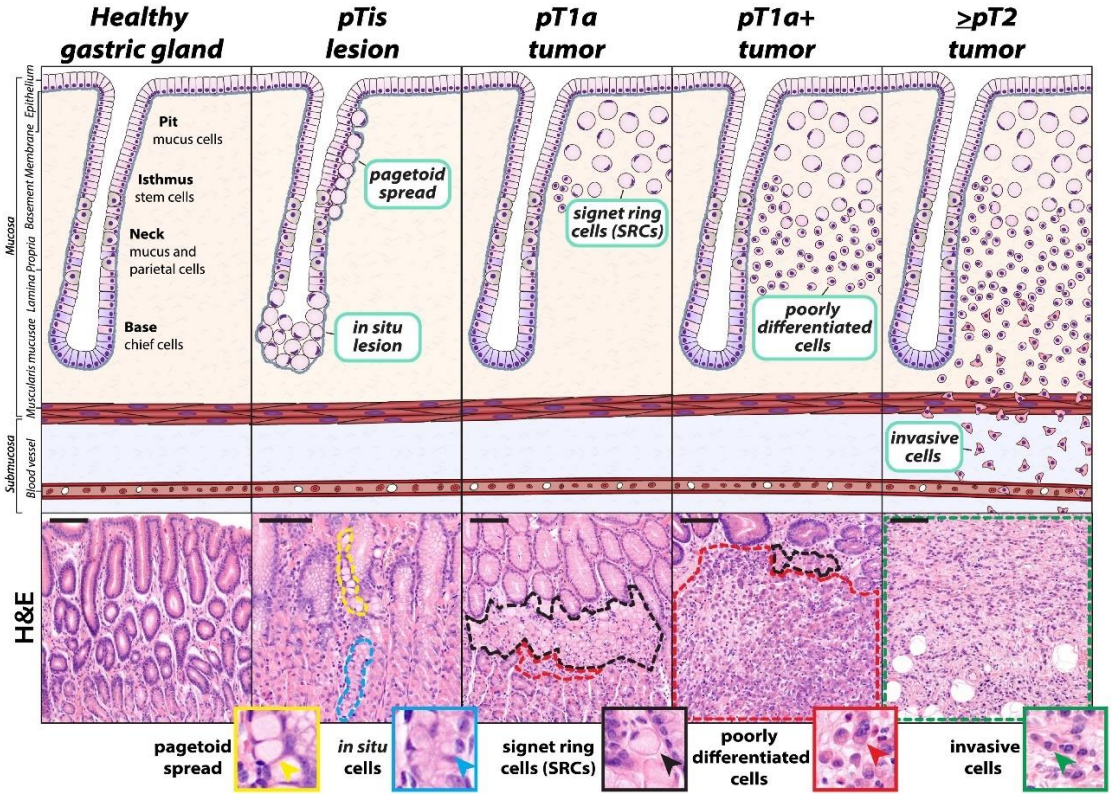


Figure 1. Histological classification of Diffuse Gastric Cancer (DGC) stages. Premalignant stages (pTis) comprising *in situ* lesions (H&E staining blue outline) and pagetoid spread (yellow). Early cancer stages (pT1a, pT1a+), comprising SRCs and poorly differentiated cells invading the lamina mucosa, outlined in black and red respectively. And advanced cancer stage (≥pT2), where the tumour invades the submucosa and muscle layers, outlined in green. *Adapted from* **Monster** *et al*, **2022**[35].

Different from other gastric adenocarcinomas, DGC tends to metastasize more within the peritoneum or to bone[41].

Some clinical features have been associated with patients with HDGC, such as the presence of chronic gastritis in the gastric mucosa. However, unlike IGC, *Helicobacter pylori* infection and intestinal metaplasia are not commonly found in association with this syndrome[36,39,42]. Nonetheless, it is recommended that in case of presence, *H. pylori* infection should be eradicated[30,43].

1.1.2. Lobular Breast Cancer (LBC)

Lobular breast cancer (LBC) is one of the two main types of breast cancer (BC), representing approximately 15% of the disease diagnosis, a number that has been increasing in the recent deceades[44–46]. However, LBC is still a rare cancer in hereditary context, when compared to sporadic LBC. It is associated with the loss of function of E-cadherin, encoded by *CDH1*, thus it is possible to consider LBC another HDGC-related cancer[9,30].

LBC is histologically different from other types of breast cancer, and more difficult to detect. Tumour cells originate in the lobular tissue of the breast, infiltrating the surrounding tissue usually in a single file of cells or in small clusters, with a more diffuse distribution and not affecting the organ architecture too much[47,48]. It is also associated with a higher risk of affected women to develop bilateral BC[49].

TNM classification is also the current in use for breast cancer, and LBC as an extent. Regarding the tumour (T) category, it is classified basically regarding the size of the tumour. Carcinoma *in situ* findings are classified "Tis", specifically for lobular breast cancer it can also be identified as LCIS. T1mi, -a, -b, and -c are classifications for tumours from below 0.1cm (T1mi) to 2cm (T1c), T2 identifies tumours from 2 to 5cm, and T3 larger than 5cm[33]. T4 is the only classification in this category that specifies not size, but the extent of the tumour to chest wall or tissue, including inflammatory breast cancer. The N category classifies the extent of cancer to axillary (N1), breastbone (N2) or below the collarbone and internal mammary nodes (N3), while the M category informs of the existence of distance metastasis (to other organs) or not[33]. Some findings can also be risk factors for invasive breast cancer, such as the non-invasive, abnormal proliferation of cells within breast lobes, lobular intraepithelial neoplasia (LIN), and should not be disregarded during surveillance procedures[50].

1.1.3. Clinical Criteria for HDGC Screening

Genetic screening for *CDH1* germline variants is currently applied when there is a suspicion of HDGC within a family, according to criteria established in 1999, with the latest revision in 2020 by the International Gastric Linkage Consortium (IGCLC)[30,51]. That is, to be tested, a patient must meet at least one of the following: family or individual criteria.

Family criteria include: i) the presence of two or more cases of gastric cancer in the family, with at least one being DGC; ii) one or more cases of DGC and LBC in different family members; or iii) more than 2 LBC cases in family members below 50 years of age[30].

As for individual criteria, a possible HDGC proband must have i) DGC before the age of 50 years old, ii) any age of Māori descendance, iii) DGC and family history of cleft lip/palate; iv) DGC and LBC both before 70 years old; v) bilateral LBC before 70; or vi) *in situ* or pagetoid spread signet ring cells in stomach epithelium before 50 years old[30].

Tests are performed in individuals from legal age of consent meeting these criteria and, when possible, a multigenerational family pedigree is advisable. Younger family members may be considered for screening according to family history[30].

Individuals are tested primarily for *CDH1* and *CTNNA1* variants, with focus on pathogenic and likely-pathogenic variants interpreted according to ACMG/AMP guidelines[30]. Since individuals carrying any of these variants are more susceptible to HDGC development, they start to be followed by a team of health professionals for surveillance.

Index patients usually show advanced gastric cancer at a young age, and further study of family history may reveal cases of DGC or LBC in other members. Many of these cases still have an unknown genetic cause for HDGC after *CDH1*/*CTNNA1* screening, with only around 40% of families presenting a known *CDH1* pathogenic or likely-pathogenic variant[27,28].

In 2023, Garcia-Pelaez *et al* showed that some families with LBC who did not meet the HDGC criteria established in 2020 displayed pathogenic and likely-pathogenic *CDH1* variants[52]. These findings suggest the need for future reformulation of HDGC criteria to include a broader family and individual LBC criteria. In addition, this study reclassified six LBC-related *CDH1* variants as benign, likely-benign, or uncertain significance variants. This highlights the need for careful and critical considerations when performing variant classification and their actionability for a thorough clinical procedure's application.

1.1.4. HDGC Surveillance and Treatment

After HDGC genetic criteria is met, either familial or individual, the individual is subjected to an established management procedure (Figure 2). If the patient is positive for any variant of uncertain significance, is advised to stay under surveillance with considerable periods of interval (2+ years). In case of a pathogenic *CDH1* or *CTNNA1* variant being detected and there are DGC cases in the family, risk reducing gastrectomy (RRG) is recommended, specially before the age of 30, due to the high risk and low efficacy of other surveillance methods to detect early stages of DGC[30]. This approach is not recommendable after 70 years old, due to the risk associated with the procedure.

Patients can opt to not perform RRG, in those cases patients are advised to monitoring by regular endoscopy with multiple random biopsies (annual), and, when present, eradication of *Helicobacter pylori* is advisable. In addition, *CTNNA1* variant carriers that remain asymptomatic are recommended to follow the same surveillance methods, with the consideration of a possible RRG[30].

Breast magnetic resonance imaging (MRI) is the best surveillance LBC method for women from HDCG families. MRI is still an effective method to identify possible early manifestations of tumours, further validated or not by biopsy. Clinicians may recommend bilateral risk reducing mastectomy (BRRM) to patients, most of the times depending on family history and cancer incidence within the pedigree[30].

All the options should be considered with the utmost caution by a multidisciplinary team of clinicians and the patient, as risk reducing measures are usually invasive, with consequences for the rest of the patient's life, without guarantee that without it the disease would progress during their lives.

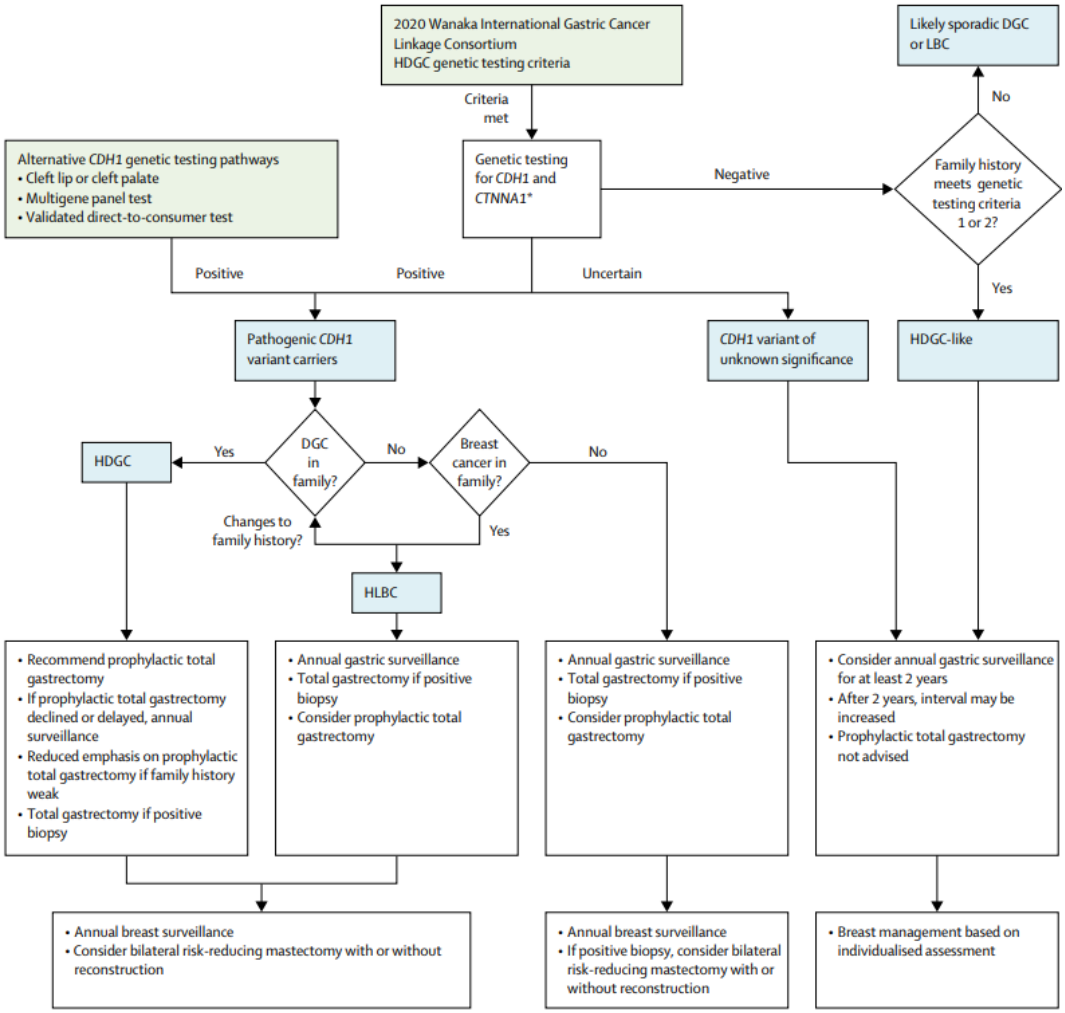


Figure 2. Schematic representation of the pathways of management and surveillance of individuals from HDGC families. These patients either meet the criteria for HDGC genetic testing or were found to be carriers of *CDH1* pathogenic variants through other source. *Adapted from* **Blair** *et al*, **2020**[30].

1.1.5. HDGC-related Cancer Lifetime Risk and Epidemiology

HDGC is currently estimated to have an incidence between 5 and 10 individuals per 100 000 births[8,53]. Risk estimations for advanced DGC is still variable among studies, with an incomplete penetrance of the disease in HDCG families with *CDH1* germline variants[30,54,55]. It has an average

age of onset below 40 years old and lifetime risk estimations have been varying between 40% and 67% in men and 63% to 85% in women for DGC, and a range from 39% to 52% for women in LBC[13,54,56–58]. There may be possible overestimations due to the sample pool and how the families were ascertained[59]. In latest studies, following a more restrictive IGCLC criteria from 2015, it was reported DGC lifetime risk estimations of 37% to 42% in males and 24% to 33% for females, while values for LBC were set at 55%[59,60]. DGC lifetime and cumulative risks vary substantially between studies and according to stricter or loosened guidelines, or the size and structure of the study group[30,59].

1.1.6. *CDH1* Founder Variants – The c.1901C>T Variant

The focus of this thesis will be HDGC families that share a single pathogenic germline variant between them, the *CDH1* c.1901C>T variant. It is located in the exon 12 of the *CDH1* gene (Chr16: 68822190, GRCh38) and consists in the substitution of a cytosine (C) to a thymine (T) in the 1901 position of the gene. It was classified as a missense variant, with the alteration of an Alanine (A) to a Valine (V) at the codon 634.

This variant was first reported in 2002 on cell lines from human colon carcinoma[61]. It was identified for the first time in a Portuguese proband with early onset DGC in 2003[14], and reported in a family with two members with DGC, fulfilling HDGC criteria, in 2004[62]. Until 2021, eight more families were identified[63,64] with HDGC-cancer affected members, with four sharing a common ancestor through haplotype analysis[63]. Thus, c.1901C>T was proven to be the first *CDH1* founder variant in a region of Portugal[63,65].

Barbosa-Matos *et al* also confirmed that the c.1901C>T variant promoted premature truncation triggered by cryptic splicing[63], which supports the pathogenic classification according to ACMG/AMP guidelines[66].

1.2. Lifetime Risk Estimations of Genetic Variants in Hereditary Cancer

Lifetime risk estimations are a useful resource to produce information about the likelihood that an individual with a particular trait has to develop a condition at some point during their lifespan. This is used abundantly in cancer research[67,68]. Hereditary cancer is thus a suitable field to apply these estimations.

It can be said that the first study of cancer segregation within a family was in 1866 when Pierre Paul Broca started to describe his wife's family breast cancer's history, with a group of ten affected women across three generations[69]. Fast forward more than 100 years and many genes were starting to be uncovered as possible causes for the development of cancer or cancer-related syndromes

such as *TP53*[70], *BRCA1/2* for breast and ovarian cancer[71], or genetic cancer syndromes related to hereditary colon cancer[72].

In the last two decades, the estimations for lifetime risk and penetrance assessment have been developing and are now a firmly rooted in cancer genetics research.

1.2.1. Lifetime Risk Estimations for *CDH1* Germline Variant Carriers

In 1999, Guilford *et al* reported for the first time a set of germline variants in the *CDH1* gene that were associated with familial gastric cancer[13]. Resorting to the family pedigrees, it was possible to observe that the segregation of cancer across the family was of dominant inheritance of genetic susceptibility, but incomplete penetrance.

In a 2001 study, Pharoah *et al* included 11 families with 11 different *CDH1* nonsense variants for penetrance estimation[57]. Selection criteria was based on Caldas *et al*, 1999[56], and families were considered if there were three or more cases of diffuse gastric cancer and at least one positive tested individual for a *CDH1* nonsense variant associated with disease. These families included 476 individuals, with 80 gastric cancer cases with an average age of onset of 40 years, with 44 cases with histological confirmation of DGC, and seven cases of breast cancer (mean age: 53), two confirmed LBC. The cumulative risk of gastric cancer was 67% for men and 83% for women at 80 years of age, with the cumulative risk for breast cancer in women was 39%[57]. There may have been the problem of overestimation in these results due to ascertainment bias or another gene-interacting external factor.

In 2007, Kaurah *et al* estimated the cumulative risk of gastric cancer for 4 families carrying a *CDH1* deletion to be 40% (95% confidence interval [CI] of 12%-91%) for males and 63% (95% CI of 19%-99%) for females by age of 75 years. For breast cancer, the cumulative risk at 75 years was calculated to be 52% (95% CI of 29%-94%) [54]. The method applied for analysis was the same as used by Pharoah *et al* in the previous study[57], using a conditional likelihood method applied in MENDEL software[73], where this likelihood was maximised with the phenotype of the family at ascertainment and the genotype of the index case. Parameters consisted in log relative risks of gastric and breast cancer based on the general population incidence data. Posterior cumulative risk per gender was extracted from these calculations. Only patients with clinically diagnosed gastric and breast cancer were considered.

Hansford *et al* used the information 75 families with reported *CDH1* variants to perform penetrance analysis, resorting to the MENDEL program, with log relative risks of both gastric and breast cancer based on general population risk, with assumed rare mutation frequency (0.001)[58,73]. For breast cancer, the relative risk was considered constant with age, which could be a limiting factor of this study. Obtained cumulative risk of gastric cancer was 70% (95% CI of 59%-80%) for men and 56% (95% CI of 44%-69%) for women by age of 80 years. In addition, breast cancer cumulative risk in women by 80 years was 42% (95% CI of 23%-68%)[58].

In 2019, Roberts *et al* designed a penetrance study for *CDH1* pathogenic variants with the goal to diminish the ascertainment bias for the high cumulative risk values presented in the previously

mentioned penetrance studies[59]. This analysis included 41 families with a complete pedigree from a cohort of 75. Only 25 of these families met the 2015 IGCLC[74] criteria in use at the time of the study. Surveillance, Epidemiology, and End Results (SEER) Program[75] was used for the incidence of general population inference. Standard Incidence Ratio (SIR) was used to obtain cumulative risk estimations and the number of individuals with cancer was modelled resorting to a Poisson binomial distribution[59]. Cumulative risk for gastric cancer was 42% (95% CI, 30%-56%) for men and 33% (95% CI, 21%-43%) for women by age 80 years. For breast cancer in women, the cumulative risk was estimated at 55% (95% CI, 39%-68%)[59]. This study bases the cumulative risk estimations on the SIR-SEER incidence ratio, assuming a proportional incidence of disease.

Penetrance estimations for *CDH1* pathogenic variants highlight some limitations that are transversal to practically all studies. Ascertain bias is thoroughly present in these studies, leading to possible overestimation. Besides, multiple pathogenic variants are considered in each penetrance analysis, diluting the true effect penetrance of each variant. This highlights the need for penetrance estimations in each variant context.

1.2.2. Lifetime Risk Estimation Studies for *CTNNA1* Germline Variant Carriers

Recently in 2022, Coudert *et al* estimated for the first time the risk for DGC of carriers with germline pathogenic variants in the *CTNNA1* gene[76]. In this study, they applied the genotype restricted likelihood (GRL) method to obtain penetrance estimations[77]. This method considers the genotype of tested family relatives, conditioned by phenotypes observed and the index's carrier status to calculate its estimations. Two models were used, a parametric function – Weibull – and a 3-points nonparametric (NP) function. This study included 13 *CTNNA1* families with gastric cancer, with 46 carriers of 12 different pathogenic or likely-pathogenic *CTNNA1* variants. Due to lack of genotype information in some families, only eight were considered for risk analysis. Cumulative risks were 49% with the Weibull and 57% with NP methods, by age 80[76]. It was also reported that risk ratios to general population were higher at early ages, being in line with the early onset of DGC in hereditary settings[8,30,78].

An important characteristic of this study is the inclusion of pT1a lesions found in histopathological reports as additional cancer events for the risk estimations. It is worth noticing that, however innovative, this study, as the others above, consisted of penetrance studies for several different pathogenic variants from a gene, which may also have its limitations.

1.3. The Transcriptome of HDGC Cancers and Pre-malignant Lesions

The standard histopathological system of patients' tissue analysis often fails to recognize the molecular heterogeneity of gastric cancer. Cancer research has been revolutionized by the emergence of Next Generation Sequencing (NGS) which comprises, among other tools, RNA-sequencing[79,80]. This high-throughput technique can help to improve molecular knowledge on carcinogenesis[81], complement immunohistochemistry reports in search for cancer biomarkers[82], reveal regulatory regions important to gene regulation in cancer context[83], of identify novel targets for cancer therapy[84,85]. Through this technology, profiles of differentially expressed transcripts in normal/tumour tissue of GC can be drawn and, ultimately, potential markers of malignancy can be identified. These can be integrated with the clinical and pathological characterization of patients to provide a more informed decision of diagnosis and prognosis.

In the literature several transcriptomic analysis reports can be found for GC. Tong *et al* proposed that aberrant expression of kinase genes in gastric cancer could pose as triggers for gastric cancer progression, invasion and metastasis, posing as possible biomarkers for the disease[86].

An interesting finding in Carino *et al* report from 2021, showed several differentially expressed genes for both diffuse and intestinal gastric cancer, with only partial overlap between samples of both cancer types. It was reported a significant enrichment in pathways involved in cell division and adhesion, methylation and lipid metabolism. Additionally, PI3K-AKT and CXCL1-CXCR2 pathways were revealed as possible targets for DGC[87]. These are signalling pathways involved in cell growth, proliferation and angiogenesis (PI3K-AKT), and regulation of chemotaxis and inflammation (CXCL1-CXCR2). These are important finding due to the dysregulation of these pathways could result in cancer growth, invasiveness, and metastasis[87].

Transcriptomic landscape of cancer is still an evolving field and novel techniques keep surginging every year.

2. Rational and Aims

The work of this thesis is divided into two main aims. As first goal, we intend to establish a robust clinical and demographic database of 11 known families with HDGC from the Northern region of Portugal. These families share the same pathogenic c.1901C>T *CDH1* founder variant across different generations, with known cases of diffuse gastric cancer and/or lobular breast cancer. This database will pose as the basis for a first-time study of intrafamilial lifetime risk estimations of HDGC among carriers of a single known pathogenic variant. We aim to find an explanation for the low penetrance of disease among carriers of this variant, while at the same time understand why most of the affected patients manifest the disease at an early age and with an aggressive phenotype, usually with a poor prognosis.

Secondly, we will study the transcriptomic landscape of more than 20 thousand genes of selected cases with different DGC and LBC stages to find genes that could be differentially expressed among normal, tumour and metastatic tissues. We aim to uncover genes and transcriptomic programs that could be essential for tumour development and proliferation and become targets for possible therapeutics or markers in surveillance genetic testing.

We plan to answer these questions in a logical step-by-step approach through the following specific aims:

- Establish a robust database with data of different clinical features and manifestations from individuals belonging to c.1901C>T variant carrier families.
- Generate extended pedigrees for 11 c.1901C>T variant carrier families to better understand variant and disease segregation.
- Study clinical features shared among carriers and the surveillance procedures applied to patients with or without manifestation of HDGC disease from the moment of testing until last news reporting.
- Estimate the disease lifetime risk in c.1901C>T variant carrier families.
- Disclose the transcriptional profile of *CDH1* c.1901C>T-related DGC and LBC.

With this work we intend to develop a unique and robust clinical database of HDGC patients to improve data access and serve as the foundation for clinical research and management. We aim to uncover intrafamilial lifetime risk estimations for HDGC that could be helpful in establishing more precise management and surveillance guidelines in the future, which are still leaning towards a more invasive approach.

3. Material and Methods

3.1. The FCT funded project LEGOH

This master's thesis work is included in the LEGOH project (PTDC/BTM-TEC/6706/2020), with the title "Life-time Risk Estimations and Genetic Modifiers of Hereditary Diffuse Gastric Cancer". Its main goal is to, for the first time, generate intrafamilial lifetime risk estimations for a single *CDH1* variant in HDGC setting, and establish a cancer predisposing molecular profile to impact early disease detection and HDGC management within affected families.

The pathogenic founder variant in focus in this study is the *CDH1* c.1901C>T variant. It was first reported in Portugal in 2003, related to early onset gastric cancer probands (cite Suriano, 2003 Hum Mol Genet).

Every data collected is under the scrutiny of the project PI, Professor Carla Oliveira PhD, and its dedicated team of researchers and clinicians.

3.2. Study Cohort

This study cohort is comprised of individuals from different HDGC families related to 11 proband cases. These were individuals either tested at *Centro Hospitalar e Universitário de São João* (CHUSJ) or Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), while some were collected with CHUSJ patients' hearsay or reports from other institutions of known relatives of each family. Blood samples were collected from individuals participating in genetic counselling appointments to be used for genetic screening of the c.1901C>T variant.

Usage of all data provided and analysed in this master's thesis work is under CHUSJ ethics committee approval and with signed consent of all living participants or under clinical investigation law (21/2014). All individuals participated in genetic counselling appointments and the majority of living c.1901C>T carriers is still under surveillance at CHUSJ.

All the demographic and clinical data was obtained through a thorough research of patient information in several CHUSJ sources. These were clinical appointment diaries, genetic testing reports, electronic health records of each patient, histopathology reports from biopsies, gastrectomy and mastectomy procedures. All information was gathered with the supervision and guidance of clinicians involved with these patients' care and surveillance, and with full discretion regarding ethics and data protection.

For the extended database, all data retrieved from patient hearsay about family history was taken into account with protection of confidential data and with full approval of all parts. Gathered information can be observed in future penetrance and lifetime risk estimations analysis with complete anonymity of the cohort.

The cohort was divided into 11 families according to the proband of each one. Identification was attributed according to follow up of these families at CHUSJ, from F1 to F11. Families that were posteriorly found to share proximal ancestry or intermarriage with resulting offspring were merged and considered as a single family.

3.3. Generating a clinical and demographic database and extended pedigrees of HDGC families

3.3.1. The extended HDGC cohort families and pedigrees

Development of extended pedigrees was enacted with a thorough and systematic approach. The focus was to gather information about known carrier status, DGC and LBC events and respective age of onset, and age of last news regarding aforementioned information.

Individuals tested and followed at CHUSJ had signed consent for full disclosure of needed data either from clinical records or histopathological reports. For the extended pedigree cases, this information was provided by CHUSJ patients with full anonymity and, whenever possible, doublechecked with clinicians from other institutions following individuals from the same families.

Pedigrees were drawn to verify the c.1901C>T variant segregation across these families. Depending on family size and available data, at least three generations of each family were traced back to better acknowledge the effect of this segregation. Unless tested at CHUSJ, branches of families descending from known noncarriers of the c.1901C>T variant were removed to avoid population noise.

Pedigrees were developed resorting to PhenoTips software (https://phenotips.com/) for an easier handling of genomic and health records data collected from each individual in the study.

In addition to familial kinship, DGC and LBC (in females) events and age of onset was collected, when possible, from aforementioned patients' hearsay, older CHUSJ clinical records and from other medical institutions. Gathered information was used for intrafamilial lifetime risk analysis for the *CDH1* c.1901C>T variant.

3.3.2. The CHUSJ genetic screening cohort

With the partnership of a clinical team from CHUSJ, all pertinent available data was collected from the genetically tested and/or under surveillance individuals at this institution. The data to be extracted was stratified according to the goals of the study: development of intrafamilial lifetime risk analysis for the *CDH1* c.1901C>T variant and genomic and transcriptomic landscape studies across HDGC families.

According to these goals, the types of data collected were subdivided into five categories and further detailed in Table 2:

- Demographic data: year of birth/death, age, pedigree information.
- Clinical data: carrier status, carriers follow-up, nutritional data.
- DGC related data: presence/stage of disease, histology, surveillance, prophylactic and treatment measures applied.
- LBC related data: presence/stage of disease, histology, surveillance, prophylactic and treatment measures applied.
- Other cancer manifestation data: type, age of onset, histology.

Table 2: Description of attributes used to build demographic and clinical database.

Database attribute entry	**Description of information provided**
Family ID	Family identification tag
LEGOH ID	Individual identification in LEGOH project
Individual General Database ID	Individual identification tag
Gender	Gender of the individual
Mother ID	Family Member ID of the mother of the individual
Father ID	Family Member ID of the father of the individual
Carrier Status	Carrier status of the CDH1 c.c.1901C>T variant
Proband	Proband/index case of a HDGC family
Family Pedigree Generations	Generation of the individual in the family pedigree
Year of Birth	Year of Birth of the individual
Age in 2023	Age of the individual in 2023
Age @ Last News for Penetrance Studies - DGC	Age at last considered event for DGC penetrance study (last appointment, prophylactic surgery, treatment, death)
Age @ Last News for Penetrance Studies - LBC	Age at last considered event for LBC penetrance study (last appointment, prophylactic surgery, treatment, death)
Age @ Last Appointment	Age at last medical appointment
Year of Last Appointment	Year of last medical appointment
Age of Death	Age of death
Age @ Molecular Study	Age of genetic testing for *CDH1* c.1901C>T variant
Molecular Study Year	Year of genetic testing
Hospital at Diagnosis	Hospital where molecular diagnosis was performed
Confidence of Carrier Status	Confidence in molecular diagnosis result (if tested or validated at CHUSJ)
Surveillance Status	Status of patient surveillance
LEGOH Classification	LEGOH project classification in groups of interest
Gastric Cancer (GC)	Diagnosis of diffuse gastric cancer
GC Year of Diagnosis	Year of diffuse gastric cancer diagnosis
GC Age Onset (Years)	Age at which diffuse gastric cancer was diagnosed
Timing of GC detection and histology	Detection event and stage of disease
GC Classification	Diffuse gastric cancer classification - premalignant, early or advanced
GC Stage	Gastric cancer stage by TNM staging system

Treatment for DGC	Information about gastric cancer therapy performed
Tumour Immunohistochemistry (Yes/No)	Analysis of tumour with specific immunohistochemistry markers
Year of Latest Gastric Biopsy	Year of the last recorded gastric biopsy
Age @ Latest Gastric Biopsy	Age at last recorded gastric biopsy
Neoplastic Lesion in Gastric Biopsy Results	Neoplastic lesion findings during gastric biopsy
Gastric Biopsy Results: Gastritis	Gastritis findings during gastric biopsy
Gastric Biopsy Results: Presence of Mucosa Abnormalities	Mucosa abnormalities findings during gastric biopsy
Presence of *Helicobacter Pylori*	*Helicobater pylori* findings during gastric biopsy
Gastrectomy	Information about performed gastrectomy
Year of Gastrectomy	Year of performed gastrectomy
Lesions in Gastrectomy	Neoplastic lesion findings during gastric gastrectomy
Type of Lesions in Gastrectomy	Detailed information of neoplastic lesion findings during gastrectomy
Gastrectomy Results: Mucosa Abnormalities	Mucosa abnormalities findings during gastric gastrectomy
Age at Gastrectomy	Age at gastrectomy procedure
Years since Gastrectomy	Years passed from gastrectomy to 2023
Gastrectomy with CHUSJ complementary report	Information about CHUSJ gastrectomy reports completion
Observations after Gastrectomy	Relevant events and findings after gastrectomy procedure until 2023
Height (cm)	Height of patient (only in those submitted to gastrectomy)
Weight (kg) - Before/After Gastrectomy	Weight of patient before and after gastrectomy procedure
BMI before Gastrectomy	BMI before gastrectomy
BMI after Gastrectomy	BMI after gastrectomy
Latest Nutritionist follow up	Year of last nutritional medicine appointment
Breast Cancer (BC)	Diagnosis of lobular breast cancer
BC Age Onset (Years)	Age at which lobular breast cancer was diagnosed
Timing of BC detection and histology	Detection event and stage of disease
BC classification	Lobular breast cancer classification - premalignant or advanced
BC Stage	Breast cancer stage by Nottingham and TNM staging systems
Treatment for BC	Information about breast cancer therapy performed
BC Unilateral or Bilateral	Information regarding lobular breast cancer affected breast
Year of latest Breast Biopsy	Year of the last recorded breast biopsy
Breast Biopsy Results	Neoplastic lesion findings during breast biopsy
Mastectomy Right (Yes/No)	Information of right breast mastectomy
Mastectomy Left (Yes/No)	Information of left breast mastectomy
Mastectomy Unilateral or Bilateral	Information of which breast was removed
Year of Mastectomy	Year of performed mastectomy
Lesions in Mastectomy	Neoplastic lesion findings during gastric gastrectomy
Type of Lesions in Mastectomy	Detailed information of neoplastic lesion findings during gastrectomy
Age at Mastectomy	Age at mastectomy procedure
Other Cancers	Information of other reported cancers
Year of Other Cancer Diagnosis	Year of other cancer diagnosis
Other Cancers Age of Onset	Age at which other cancers were diagnosed
Other Cancers (Histology)	Detailed information of neoplastic lesion of other cancer types

This data was essential for further studies about surveillance methods applied and their efficiency, as well as statistical analysis of pathways-of-care of and look for correlations between clinical features and DGC or LBC development.

3.3.3. Criteria for Carrier Status and Cancer Onset

For the database comprised of patients tested and followed at CHUSJ, criteria for carrier status were based solely on genetic testing, with the presence or not of the c.1901C>T variant.

When considering the extended families, carrier status was judged on two different accounts: either the carrier status was confirmed by genetic screening at another institution, or the individual had to be an obligated carrier (OC). For an individual to be classified as an OC, it had to meet one of the following criteria:

- Be both the offspring and parent of variant carriers;
- Be both the parent and sibling of variant carriers and HDGC affected individuals.

Regarding the verification of an individual with a cancer event, these were divided into histologically confirmed DGC and/or LBC – by CHUSJ electronic heath records and histopathology reports – and unknown cancer histology.

Individuals with unknown cancer histology were considered affected for lifetime risk analysis if they were reported as displaying either gastric of breast cancer, and were at least second-degree relatives with a confirmed variant carrier.

Selection of non-tested individuals for risk estimations was based on either:

- Individuals with confirmed carrier status at another institution or previously classified OC;
- Direct descendants of c.1901C>T variant carriers;
- Mates of variant carriers with offspring;
- Siblings of variant carriers;
- Cancer affected individuals with at least second-degree relation to variant carriers.

3.3.4. Statistical and surveillance analysis of CHUSJ HDGC patients' database

After establishing the HDGC clinical database, basic statistics were used to describe the information gathered from patients tested and under surveillance at CHUSJ. General frequency and penetrance of DGC and LBC cases were calculated considering every known case of cancer across the database, and all c.1901C>T carriers, respectively.

Surveillance at CHUSJ status was considered when patients' health records – either biopsy, gastrectomy, and/or mastectomy – were found at CHUSJ database, implying that the patient underwent surveillance procedures at some point at the institution. Considering this assumption,

patients that are no longer under follow up, either by death, loss of contact, changing of institution or due to old age, are considered under surveillance until the age of last news.

Frequency of different surveillance and preventive measures were calculated based on carriers that were currently or previously under follow-up at CHUSJ.

Fisher's exact tests were performed to find possible statistical significance between clinical features, outcomes and/or preventive measures of these patients. Median tests were used in some clinical features analysis such as age at genetic test, risk reducing gastrectomy (RRG) and gastric cancer event, to find differences between male and female.

3.4. Lifetime risk estimations for the *CDH1* c.1901C>T variant in a Northern Portuguese cohort

Only cases meeting criteria described in 3.3.1. and 3.3.3. were considered for lifetime risk estimations. Individuals reported that were descendants from known c.1901C>T noncarriers, had no information about cancer event and its age of onset, or lacked any information about age at which information was collected were excluded from further analysis. These estimations inferred how the variant may be segregating within the familial population taking into account all known cases.

3.4.1. Cumulative incidence of DGC and LBC in Northern Portugal

DGC and LBC cumulative incidence across the Northern Portuguese population was inferred resorting to data from European Cancer Information System (ECIS) provided by the European Commission (https://ecis.jrc.ec.europa.eu/index.php).

The data collected regarded gastric cancer and breast cancer incidence between the years 2000 and 2010. This information was subdivided in values for each sex – in gastric cancer – and in age intervals of 5 years, ranging from birth until 85+ years. Based on literature and HDGC-related cancers statistics, 30% of gastric cancer incidence was considered for DGC, while for LBC it was attributed 10% of breast cancer incidence.

An incidence curve was achieved using an in-house R-based function and the *ggplot2* package (https://ggplot2.tidyverse.org/).

3.4.2. Cumulative and relative risk of DGC and LBC in the HDGC cohort

Kaplan-Meier risk curves were used to depict cumulative risk estimates of DGC and LBC for the HDGC families in this study. Data was stratified in three different groups: variant carriers, noncarriers, and unknown genotype.

For DGC, two distinct lifetime risk estimations were performed. One which considered pT1a lesions found in gastrectomy tissue as a cancer event, and another which considered only advanced (pT2 or higher) cancer as event. In the LBC analysis only pT1 or higher was characterized as events for the analysis, considering any female with *in situ* lesion as healthy.

Cases with unknown histology for both DGC and LBC risk estimates were considered as events if they met the criteria described in 3.3.3.

The Kaplan-Meier curves describe the cumulative risk of DGC (or LBC) in carriers from birth to 80+ years old in intervals of 5 years. Superior and inferior 95% confidence intervals (CI) were also calculated.

Log-rank tests were performed in the Kaplan-Meier curves, comparing groups with cancer events, to assess if there was statistically significant differences between different estimations.

Standardized Incidence Ratios (SIR) with 95% CI for both DGC and LBC were calculated from observed cancer cases divided by expected cancer cases, to estimate the incidence of DGC and LBC within the HDGC cohort and general population of Northern Portugal, resorting to the following age intervals: [0,35], (35,50], and (50,100]. These intervals were chosen to provide a better distribution of cases and stability to SIR calculations.

3.4.3. The Genotype-Restricted Likelihood (GRL) method

The GRL method was used for a robust penetrance analysis of the c.1901C>T variant in intrafamilial HDGC context. This method maximises the probability of observed genotypes from the tested family relatives at ascertainment, conditioned by phenotypes observed in the cohort and the carrier status of the index case to estimate the lifetime risk for variant carriers[77].

For possibly affected individuals, a modified parametric Weibull function was used to model risk estimations, while always considering that a portion of the cohort would never be affected. This method considers that the variant originates from only one parent in its calculations, it follows mendelian rules of variant segregation and considers the effect of variant inheritance as dominant. A non-parametric function with three fixed nodes at 35, 50 and 80 years of age was used to cross check the results obtained with the Weibull model, that does not make distribution assumptions and is better suited to small sample size. A ratio of the risk estimations between the study's cohort and the Northern Portuguese population incidence was performed as well.

Allele frequency was defined at 1/2000, given to fact that the variant is considered ultrarare in the European population, and an autosomal dominant model was used. North of Portugal overall risk for both sexes was employed for risk ratios to general population.

Similar to the Kaplan-Meier estimates, the GRL calculations were performed in three distinct analyses: GC advanced cases only as phenotypes, considering pT1a lesions as GC cases as well, and BC cases in the LBC estimations.

A sensitive analysis was used to evaluate the uncertainty of each families' impact in the GRL method. This was performed using the "leave-one-out" method, where risk curves were estimated with the absence of each family.

3.5. Selection criteria of HDGC cases and controls for different output analysis

Regarding the second part of the LEGOH project's goal, groups of interest were devised for further genomic and transcriptomic analysis.

For genomic analysis, each individual was classified according to the onset of disease and age at the time of occurrence, with focus on two distinct groups. Carriers that developed the disease before 35 years old were classified as Young Affected Carriers, while carriers that remained asymptomatic beyond 65 years of age were classified as Old Asymptomatic Carriers. However, despite being an important output analysis of the developed clinical database, genomic data analysis will not be tackled in this master's thesis.

For transcriptomic analysis, some cases of DGC and LBC, confirmed by histology, were selected. Criteria for the selection was a sufficient count of histological blocks for RNA extraction, and, whenever possible, enough material of tumour and normal tissue per sample, for pairing analysis. However, on many occasions this was not possible, especially for advanced DGC cases, due to lack of availability of normal tissue.

3.6. RNA-sequencing analysis for DGC and LBC

3.6.1. Tissue samples and parameters for RNA-sequencing

RNA was extracted from Formalin-Fixed Paraffin-Embedded (FFPE) blocks using the MagMAX™ kit (Thermo Fisher Scientific, Massachusetts, USA).

Sample selection for the extraction comprised biological material from tumoral, metastatic and normal tissue from patients followed at CHUSJ.

RNA was sequenced using the Ion AmpliSeq™ Transcriptome Human Gene Expression Kit (Thermo Fisher Scientific, Massachusetts, USA), with amplification of more than 20K genes, suitable for degraded samples, such as the ones from FFPE blocks.

3.6.2. RNA-sequencing data analysis

Search for differential expressed genes was performed in a first stage recurring to the Transcriptome Analysis Console (TAC) software from Thermo Fisher Scientific. A .CHP file per sample was used for the analysis, for probe sets and signal intensity values information.

Two analyses were performed, one for diffuse gastric cancer samples. Comparisons between normal tissue, tumour and metastasis (lymph node and distant) were performed.

The statistical method used was e-Bayes, with information of each sample's individual being accounted as a confounding variable in the analysis.

Significant log2 fold change was set at below -2 and above 2, with significant p-value being considered as <0.01. No p-value correction was considered for differentially expressed genes at each comparison, due to the low output it yielded. Principal Component Analysis (PCA) and hierarchical clustering heatmap were obtained from TAC analysis.

ClusterProfiler (v.4.4.4) R package was used to identify significantly enriched Gene Ontology functions in biological processes (BP), molecular function (MF), and cellular component (CC) categories for each analysis (FDR adjusted p-value < 0.1).

4. Results

4.1. A pioneering and comprehensive clinical database for HDGC families carrying the *CDH1* c.1901C>T variant

4.1.1. Extended database and pedigrees analyses – Cataloguing a 400 individuals' cohort with a founder genetic event

All data regarding individuals from HDGC families sharing the *CDH1* c.1901C>T variant was thoroughly examined and compiled to create a robust demographic and clinical database that will be the basis for several analysis, such as lifetime risk estimations, genomic and transcriptomic analysis.

This work highlights the importance of well collected and curated data for scientific research and health care providing.

Family pedigrees were ascertained from 11 probands carrying the *CDH1* c.1901C>T variant. From the 11 families, a total number of 400 individuals were reported, either through genetic screening at CHUSJ, older clinical records, hearsay of individuals on genetic counselling appointments, or confirmation by exterior institutions.

A first step was performed to disregard all individuals without relevant data for risk studies. According to criteria described in 3.3.1. and 3.3.3., 115 individuals without known age of last news were excluded, and 285 individuals were considered for lifetime risk analysis (Figure 3). From these, 171 had carrier status confirmation through genetic screening at CHUSJ. The c.1901C>T was found in 95 individuals – 40 female and 35 male carriers. Lastly, electronic health records from 67 patients were retrieve with information of surveillance procedures at CHUSJ.

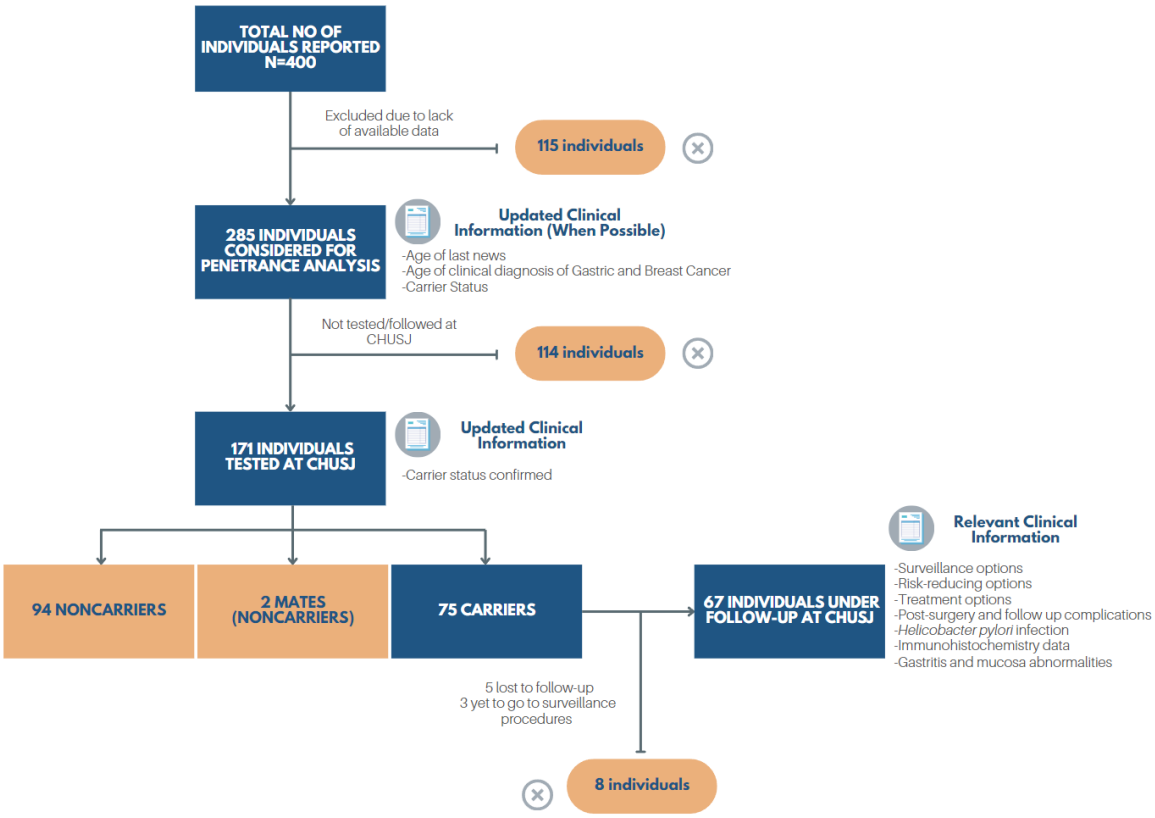


Figure 3. Diagram depiction of case selection and retrieved clinical data.

Pedigrees were drawn in PhenoTips from data provided by electronic health records or patients' hearsay. These were a tool to observe the *CDH1* variant segregation within each family, and the number of HDGC-related cancers across family generations. For data protection, these pedigrees were used for research purpose only, but a simplified representation was provided in Figure 4. The extended pedigrees also show the extent of this family study, ranging from 3 generation levels reported – families such as F6, F7, F10 and F11 (the most recent) – to families with five (F1, F2) and six (F3) generation levels and a considerable number of cases reported.

It can be observed that three of these families were merged due to shared proximal ancestry (F8 and F9) or intermarriage with offspring (F2 and F8). From now one these families will all be compiled under the F2 denomination.

Figure 4. Pedigree representation of nine HDGC families sharing the *CDH1* c.1901 variant. The nine families of the study are identified in different colours, with family generations ranging from three up to six levels.

4.1.2. Database for lifetime risk estimations – 285 individuals with known age of last news

The backbone of the database is comprised of 285 individuals from which age of last news is known. This variable is assumed from the last reported information about the individual, either genetic testing, perished by disease, last surveillance procedure, removal of targeted organs, or HDGC-related cancer diagnosis from hearsay.

The cohort is comprised of 95 known carriers of the c.1901C>T variant, 105 noncarriers and 85 individuals with unknown genotype (Figure 5). Known variant carriers present 38 cases of gastric cancer, from which 36 are confirmed DGC. Both advanced DGC (≥pT2 tumour) and early DGC (pT1a lesion) are considered as DGC events, 10 and 26 respectively. Female known carriers in this cohort also present 6 LBC cases. Four women diagnosed with LBC also presented early DGC findings.

Gastric and breast cancers were also reported in individuals with unknown genotype, from which it was impossible to retrieve the histological classification of the tumour as well. Gastric cancer was reported from 14 individuals, while breast cancer was exhibited in six females. One of the women was diagnosed with both gastric and breast cancer in her life.

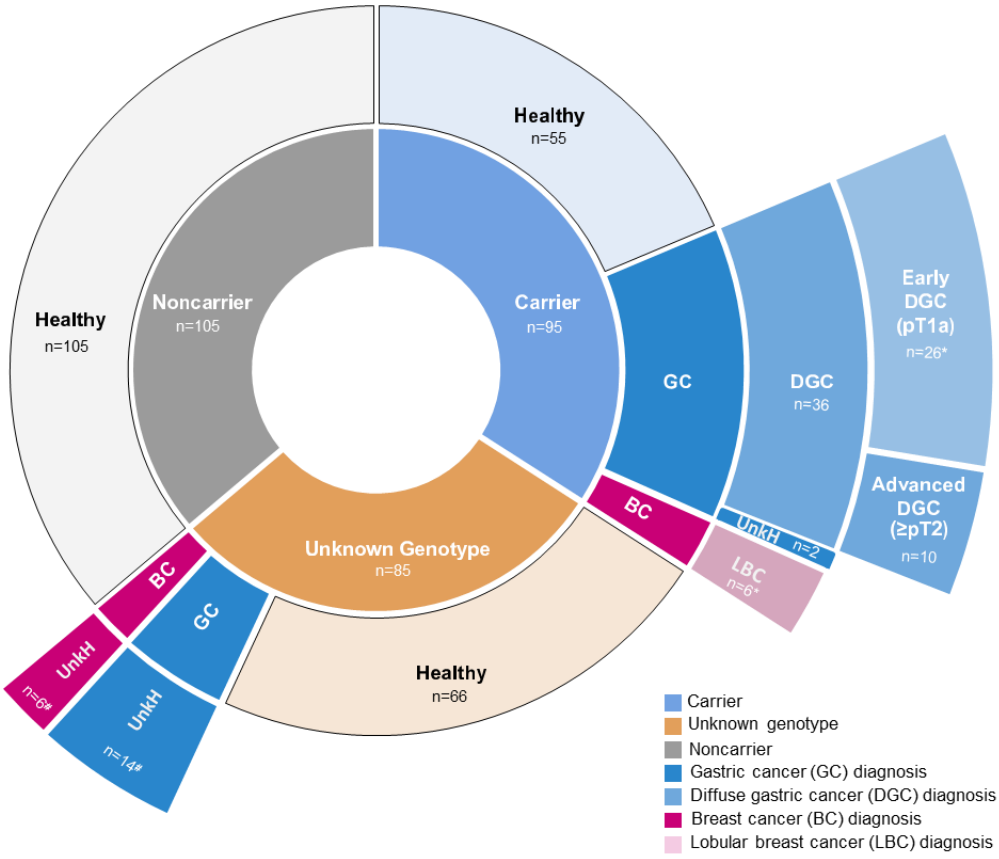


Figure 5. Chart depicting carrier status and HDGC-cancers onset in the 285 individuals' cohort. Carrier status was classified as carrier, noncarrier, or unknown genotype. Tumour histology of patients with gastric cancer (GC) and breast cancer (BC) was also attributed when proven. UnkH – Unknown histological classification of tumour (GC and BC). *Four variant carrier women had both LBC diagnosis and early DGC findings in surveillance. #One woman with unknown genotype presented both GC and BC.

Picking up from this broader database and extended pedigrees with age of last news, extensive clinical data was aggregated for patients tested at CHUSJ for posterior surveillance analysis.

4.1.3. Surveillance at CHUSJ – A gigantic pool of data, a drop of curated information

A more detailed section of the database was comprised of demographic and clinical information from the 171 individuals tested at CHUSJ. Carrier status of these individuals was confirmed after genetic testing. From these, 96 were noncarriers of the c.1901C>T variant (including two proband mates), while 75 carriers of the variant were reported, comprising the 11 probands and 64 family

members. Gender of variant carriers was split into 40 females and 35 males (Table 3). Three patients are yet to start surveillance at CHUSJ due to recent positive tests (all males), and five other patients (2 females and 3 males) were lost to follow-up before beginning surveillance procedures. Regardless, these patients had signed a consent for their clinical data to be assessed within the LEGOH project. From the 75 confirmed c.1901C>T variant carriers, 67 are or were at some point under surveillance at CHUSJ (treatment or management), with more relevant clinical data collected.

Table 3. Carrier status of individuals tested at CHUSJ.

Cases reported at CHUSJ	**Total (n = 171)**	**Female (n = 93)**	**Male (n = 78)**
Positive Carriers	75	40	35
Noncarriers	94	53	41
Mates (Noncarriers)	2	0	2

The cohort sampling was distributed among 9 different HDGC families, with families F1, F2 and F3 overrepresented compared to more recently studied families such as F10 and F11 (Table 4). Advanced DGC cases span across almost every family except F5 and F10. F5 index case was screened for *CDH1* variants for LBC onset, and F10 proband had a non HDGC cancer, but was tested due to family history of various GC events and geographical location of the family. In fact, both F10 and F11 do not follow traditional criteria for HDGC affected families but started being tested because of being from a region of Portugal with a high incidence of HDGC-related gastric cancer. The proband of F11 was clinically diagnosed with DGC abroad, before genetic testing, and requested for the family to start being followed at CHUSJ.

Table 4. Carrier status and DGC and LBC events across each HDGC family

Families Ascertained (n = 9)	**Carriers (n = 75)**	**Non-carriers (n = 94)**	**Mates (n = 2)**	**Advanced DGC (n = 10)**	**Early DGC (n = 26)**	**LBC (n = 6)**
F1	16	19	1	2	6	1
F2 (+ F8 & F9)	23	24	1	2	4	3
F3	19	19	-	2	10	1
F4	4	8	-	1	2	-
F5	5	9	-	-	3	1
F6	2	3	-	1	-	-
F7	3	7	-	1	1	-
F10	1	1	-	-	-	-
F11	2	4	-	1	-	-

Diffuse Gastric Cancer (DGC)

Clinically expressed DGC was reported in 10 (13.3%) patients (Table 4), seven were index cases (probands with manifestation of disease), one was a family member diagnosed before genetic testing, and two other cases were diagnosed in biopsy results under surveillance at CHUSJ.

From these, four of the seven probands perished due to disease, before being submitted to HDGC surveillance. Three probands and one relative were submitted to curative gastrectomy with two probands succumbing to DGC (Figure 6). From the two advanced tumour cases detected after genetic testing, one of the patients died due to DGC.

Three carriers opted to forfeit surveillance at CHUSJ due to old age.

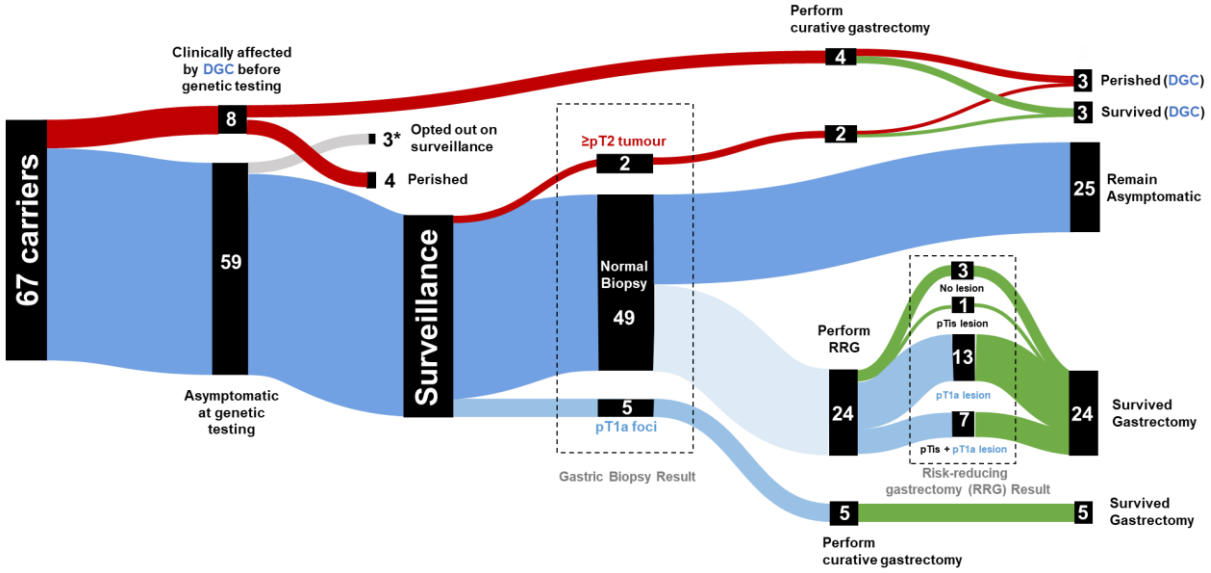


Figure 6. Chart depicting the surveillance pathways for DGC in patients from HDGC families. Surveillance or clinical events are depicted in black with information of individuals considered. Dark red represents the path of individuals clinically diagnosed with advanced DGC. Blue path englobes individuals under surveillance. Light blue represents the path of patients with pT1a findings. Green depicts successful (patient survival) curative or risk reducing gastrectomy. *Light grey corresponds to individuals that opted out on surveillance due to old age.

Early DGC foci were found in the biopsies of five patients under surveillance, which were submitted to curative gastrectomy, all surviving the procedure and the post operatory.

From the 49 carriers with normal biopsy results, 24 (49.0%) opted to perform risk reducing gastrectomy (RRG). Only three patients (12.5%) had no lesions. A total of 20 patients (83.3%) presented pT1a foci, with seven presenting *in situ* lesions or pagetoid spread. Another patient presented only these premalignant lesions in the histology reports.

Lobular Breast Cancer (LBC)

LBC only affects women, hence male carriers were removed from analysis.

Three women were lost to follow-up and three female carriers tested at CHUSJ perished due to DGC before being submitted to LBC surveillance. One patient opted out of surveillance due to old age after genetic testing.

LBC was detected in three carriers before genetic testing (4.3%), with two being family probands, F2 and F5. These all revealed different histological findings: pT2 + lobular intraepithelial neoplasia (LIN), pT1c + LIN, and unknown histology. All patients performed curative mastectomy and survived (Figure 7).

At MRI surveillance, seven of 30 women (23.3%) had possible malignant findings. After breast biopsy, two were diagnosed with LBC, two had LIN findings and three revealed normal biopsy results.

The two women with LBC detected at MRI underwent curative mastectomy, showing pT1c and pT2 tumours in addition to LIN findings. One of these women perished after mastectomy.

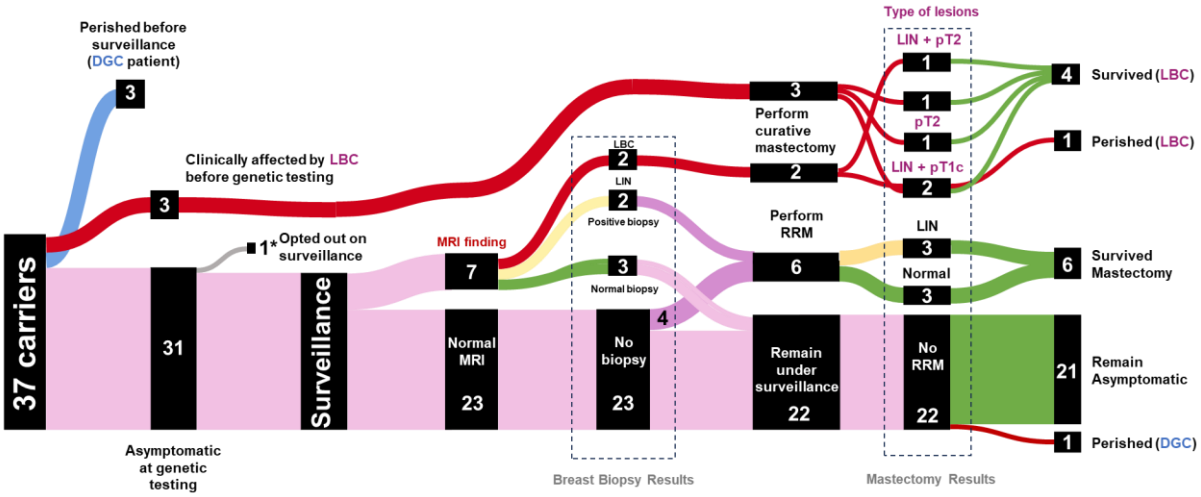


Figure 7. Chart depicting the surveillance pathways for LBC in patients from HDGC families. Surveillance or clinical events are depicted in black with information of considered individuals. Red represents women clinically diagnosed with LBC. Blue denotes women that perished due to DGC before surveillance, and dark red that perished from DGC after surveillance for LBC. Light yellow represents the path of patients with LIN findings. Pink is the path of patients that opted for risk reducing mastectomy and green depicts normal findings and women still under surveillance. *Light grey corresponds to a woman that opted out on surveillance due to old age.

Along with the two women that revealed LIN at breast biopsy, four more (17.4%) underwent risk reducing mastectomy (RRM). All six women survived RRM, with three revealed LIN findings at RRM.

The remaining 22 women were kept under surveillance procedures, however one has perished due to DGC.

Additionally, it is worth noticing that the proband of F9 (grouped into F2 family) was misdiagnosed with LBC – posterior histopathology tissue analysis showed that the patient only had lobular intraepithelial neoplasia – which for all descriptive and subsequent analysis purposes was not considered as a clinically expressed LBC case.

Clinical findings in gastric biopsy and gastrectomy reports

It was possible to collect information about histological and other clinical features from histopathological reports of tissue from biopsy or gastrectomy. Clinical records were thoroughly analysed to extract the most from the information provided at CHUSJ.

One of the most well establish findings was the presence of *Helicobacter pylori* in the carriers under surveillance. From the 67 patients, 36 (53.7%) presented *H. pylori* in the lining of the stomach, at least in one of the biopsies (Table 5). Eradication of the bacteria may occur between surveillance events.

Other retrieved information regarded the presence of gastritis signings in the stomach lining in 42 patients (62.7%) and intestinal metaplasia in 10 (14.9%), according to biopsy reports. This information sometimes does not match histopathology reports from curative or risk reducing gastrectomy. There is not a defined pattern in how the information is recorded into these reports.

It is also observable a significant lack of information in individuals where DGC was diagnosed. This may be the result of a combination of factors: no information from older probands' reports, a focus on lesions reporting and undervaluing of secondary features.

Table 5. Clinical findings in carrier reports at CHUSJ

	Patients (n = 67)	**Presence of Helicobacter pylori (n = 36)**	**Biopsy findings**		**Gastrectomy findings**		
			Gastritis (n = 42)	**Intestinal metaplasia (n = 10)**	**Gastritis (n = 21)**	**Intestinal metaplasia (n = 5)**	**Tufting (n = 4)**
DGC diagnosis before genetic testing	8	1	0	0	1	0	0
DGC diagnosis at surveillance	2	1	1	0	0	1	0
pT1a diagnosis at surveillance	5	3	4	3	0	1	1
Normal biopsy and RRG with pT1a lesions	20	14	13	2	15	1	2
Normal biopsy and RRG with no lesions	4	1	4	2	2	2	1
Normal biopsy without RRG	25	16	20	3	-	-	-
Opted out of surveillance	3	-	-	-	-	-	-

Fisher's exact test was used to find possible correlations between disease progression and neoplastic findings with clinical features such as gastritis, intestinal metaplasia, and *Helicobacter pylori*

presence. None were found in analyses either in normal biopsy or with findings. The only significant association (p-value <0.05) was found between gastritis and presence of *Helicobacter pylori*.

The reduced number of histopathological records with complete clinical feature's reports was a barrier for the statistical analysis performance. Perhaps a possible solution could be resorting to fill-in reports or checklists. This topic will be exposed in more depth in the chapter 5.

Median tests were performed to find significant differences between genders at genetic testing, RRG and disease onset. None were found.

4.2. Populational incidence and penetrance estimations – pT1a lesions as a potential incremental factor in risk estimations

To develop lifetime risk estimations for the *CDH1* c.1901C>T across the families from the cohort in study, a statistical basis was assembled for comparison. Incidence of gastric and breast cancer in the Northern Portugal population from 2000 to 2010 was collected to define general population cancer incidence. From these values, 30% and 10% was considered for DGC and LBC respectively. Incidence curves for DGC were drawn with values across age intervals of five years and separation between gender (Figure 8, panel A). It was possible to observe a slightly higher incidence of DGC in male individuals. A combination of both sexes was assembled for a general incidence curve (Figure 8, panel B).

An LBC incidence curve was performed through the same method, considering that only women develop this cancer phenotype (Figure 8, panel C).

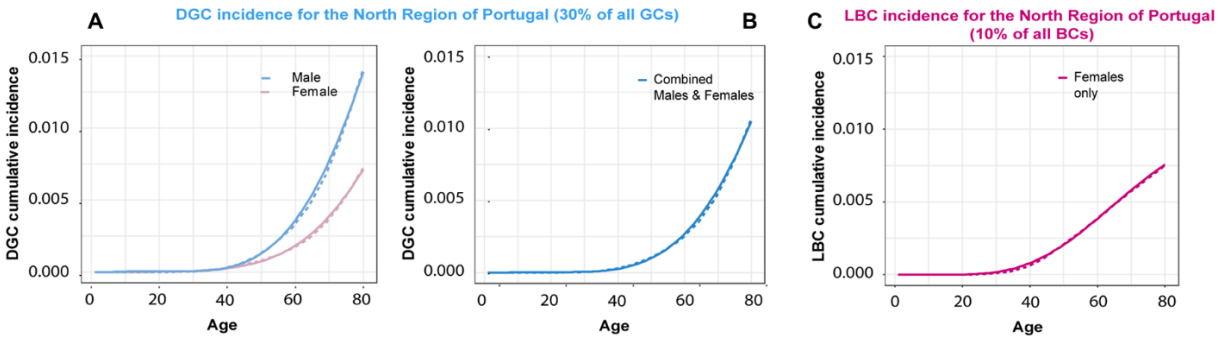


Figure 8. Incidence curves for DGC and LBC in the Northern Portugal population between 2000 and 2010. A – DGC incidence split into male and female values. B – Combined DGC incidence, depicting general North Portugal region DGC incidence. C – LBC incidence in females from the North Portugal region.

4.2.1. Kaplan-Meier and SIR analyses reveal higher HDGC-related disease risk in carriers before 50 years of age

Kaplan-Meier curves were generated to present the absolute cumulative risk of DGC and LBC development across three defined groups: c.1901C>T variant carriers, noncarriers, and individuals with unknown genotype.

Diffuse Gastric Cancer (DGC)

Two different analyses were performed. One considered both advanced and early cancer manifestations as DGC event, while another only took advanced cancer as event. For both, unknown cancer histology was considered a DGC event, if there was a reported case of gastric cancer with no histopathological confirmation.

Considering both advanced DGC and pT1a lesions as cancer events the cumulative risk of carriers is considerably higher (~60% at 65 years) than individuals with unknown genotype (~27 at 65 years). The reason behind these observations may lie on the fact that, in this cohort, pT1a lesions are only reported in individuals with known carrier status (Figure 9). These lesions are usually detected only in histological analysis of biopsies or gastrectomy tissue. Thus, access to this information was easier for patients under surveillance or previously tested at CHUSJ. Much of this data was absent from individuals with unknown carrier status.

Another finding of this Kaplan-Meier is the fact that, in known variant carriers, advanced DGC events stop showing after 55 years while pT1a lesions are reported in gastrectomy until as late as 65 years of age (Figure 9).

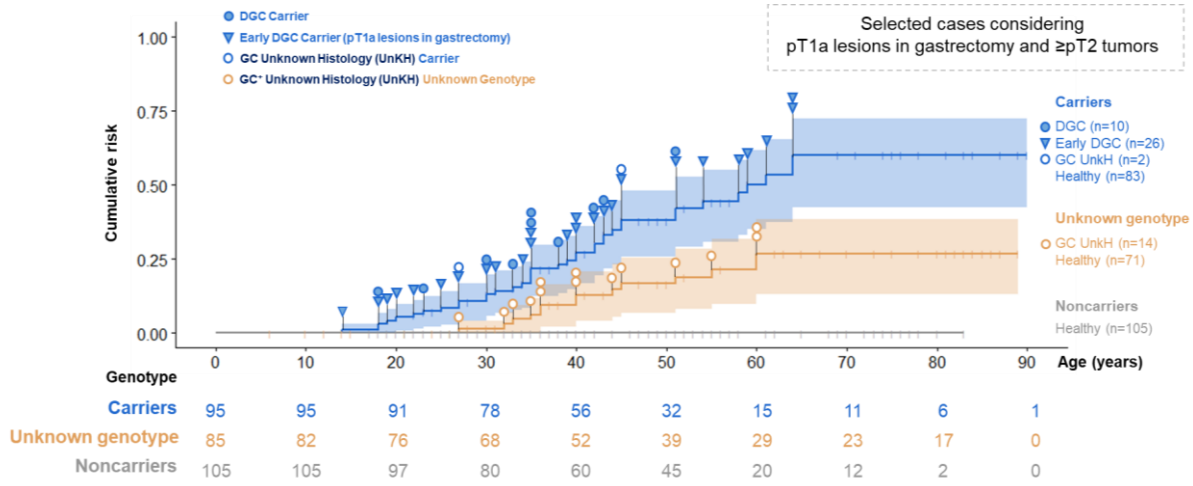


Figure 9. Kaplan-Meier cumulative risk estimations for DGC onset considering both pT1a lesions and ≥pT2 tumours as events. Curves for c.1901C>T carriers, individuals with unknown genotype and noncarriers, with a supplementary table of number of individuals observed for each group in intervals of 10 years. UnkH – Unknown histological classification of tumour.

These findings highlight a possible overestimation of DGC risk, especially after 50 and if, from this age onwards, the RRG is really the most valid option for HDGC management. This overestimation gets support from a log-rank test performed on the Kaplan-Meier risk curves (considering only carriers and unknown genotype groups), with a resultant p-value of 5 x 10^{-4}, highlighting the bigger number of events in the carrier group.

These results are corroborated when the Kaplan-Meier is generated considering only advanced DGC as event. It is possible to witness a much lower cumulative risk of DGC, at around 19% for carriers of the c.1901C>T variant at age 65, with even the risk from the unknown genotype group surpassing it from 60 onwards – approximately 27% (Figure 10). It is possible that some GC cases with unknown histology may have accounted another histologically different GC, but these were considered based on the criteria described in 3.3.3. Contrasting to the previous Kaplan-Meier, log-rank test had a p-value of 0.7, demonstrating no statistically significant difference between both groups.

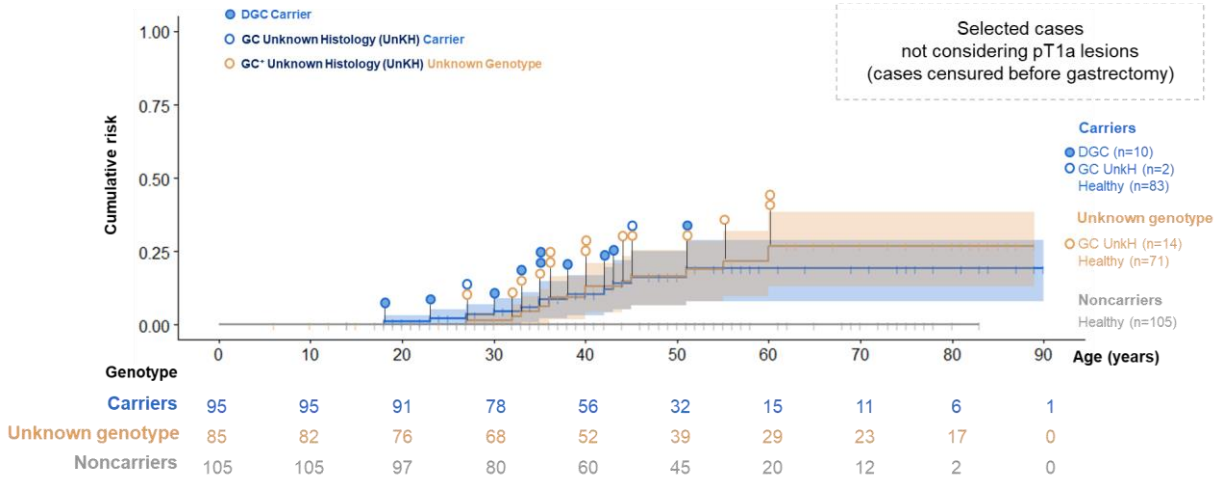


Figure 10. Kaplan-Meier cumulative risk estimations for DGC onset considering only ≥pT2 tumours as events. Curves for c.1901C>T carriers, individuals with unknown genotype and noncarriers, with a supplementary table of number of individuals observed for each group in intervals of 10 years. UnkH – Unknown histological classification of tumour.

Proceeding to relative risk calculations, standard incidence ratio (SIR) for DGC was devised. This ratio comprised the 95 known carriers from the extended pedigree. Age of last news intervals were set to three intervals: from 0 to 35, the threshold for young affected carrier's group for patients with HDGC-related cancer onset, 36-to-50, and 50 onwards. These intervals were defined based not only in groups of interest, but also to evenly distribute observations, to ensure more stability.

It can be observed that DGC carriers within these families at young age (<35 years old) have an extremely high incidence (~34 times more) of DGC compared to the general population from the same region (Table 6). Disease risk for c.1901C>T variant carriers was still superior in the age range from 36 to 50 years old, while from 50 onwards, the risk of is lower than in the normal

population, highlighting the earlier age of onset of DGC in HDGC compared to sporadic gastric cancer.

Confidence intervals (95% CI) present a wide range of values, which may be indicative of lower stability of the SIR, probably due to the low sample size of the observed cohort. Nonetheless, for the two first groups, inferior 95% CI is still above 100, meaning that, within the limits of this data, the probability of developing DGC until 50 years is higher than in the general population. In addition, superior 95% CI in the 51-to-100 interval is also below 100, showing half of the risk of general population to develop DGC, supporting that this disease tends to early onset.

In these calculations, only pT2 tumours and above were considered as DGC events, for SIR values considering pT1a lesions as well, consult Supplementary Table 1.

Table 6. Standard Incidence Ratio for DGC (only pT2)

DGC age of last news	c.1901C>T carriers (n = 95)	DGC cases* (n = 12)	SIR	SIR 95%CI inferior	SIR 95%CI superior
[0,35]	34	7	3402.04	1367.80	7009.51
(35,50]	31	4	370.88	101.05	949.59
(50,100]	30	1	9.48	0.24	52.81

Lobular Breast Cancer (LBC)

Kaplan-Meier risk curves were also estimated for LBC. This comprised only 147 female individuals. Risk curves only start at 30 and 42 years of age for unknown genotype and c.1901C>T carriers, respectively. Maximum cumulative risk for females with unknown genotype from 65 years onwards is set at 27%, and for carriers is approximately 34% (Figure 11). There was no statistically significant difference between carriers and unknown genotype groups (log-rank p-value = 0.9).

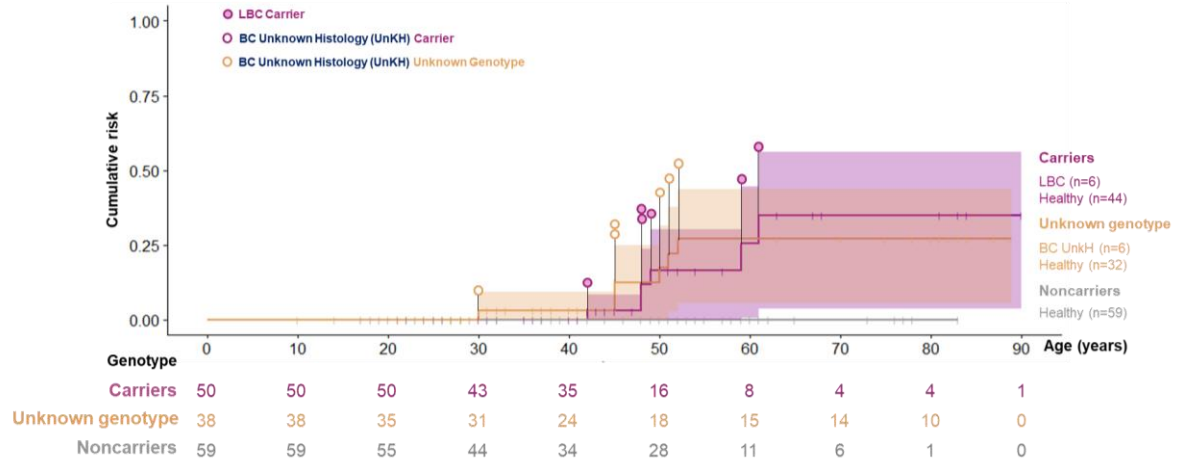


Figure 11. Kaplan-Meier cumulative risk estimations for LBC onset. Curves for c.1901C>T carriers, individuals with unknown genotype and noncarriers, with a supplementary table of number of individuals observed for each group in intervals of 10 years. UnkH – Unknown histological classification of tumour.

For LBC SIR estimations, the sample consisted of 50 female carriers from the extended pedigree database. In Table 7, higher incidence ratio of LBC in the families' cohort compared to Northern Portugal population is only observed at the 35-to-50 years of age interval. Female carriers within these families are approximately four times more at risk than women from the same age group in the region of study. It is still worth notice that the confidence intervals' range is extremely high, indicating that statistical power is not ideal and the potential existence of some bias.

Table 7. Standard Incidence Ratio for LBC

LBC age of last news	c.1901C>T female carriers (n = 50)	LBC cases (n = 6)	SIR	SIR 95%CI inferior	SIR 95%CI superior
[0,35]	11	0	0	0	6550.90
(35,50]	23	4	465.56	126.85	1192.02
(50,100]	16	2	38.26	4.63	138.22

4.2.2. GRL method highlights a higher risk of DGC between 20 and 40 years in HDGC families compared to the North Region of Portugal

Risk estimates were performed resorting to the GRL method developed by Bonaïti *et al*, with maximization of known genotypes conditioned by the observed phenotypes. For cumulative risk two methods were performed, Weibull and a 3-point non-parametric (NP).

Diffuse Gastric Cancer (DGC)

Observed in Figure 12 panel A, DGC cumulative risk in the cohort families given by this method is constant at around 17.5% for both Weibull and NP models when considering pT1a lesions alongside ≥pT2 tumours as cancer. When labelling only advanced tumour cases as cancer, there was a considerable diminishing of cumulative risk for both Weibull model (13.8% at age 65) and 3-point NP (5.2% at age 65). Both models maintain a similar increase of risk across age until 45 years old, where NP model stays constant and Weibull model continues to increase based on function assumptions.

DGC risk ratios against the Northern Portuguese population reveal a peak of relative risk between 20 and 30 years considering pT1a and ≥pT2 tumours as DGC, with both models behaving similarly (Figure 12, panel B). Risk estimations at age 30 are 2445.5 and 2466.7, decreasing to 32.3 and 32.5 at age 65, from Weibull and non-parametric models respectively. When considering only ≥pT2 tumours as cancer, the relative risk to population is much lower than the calculation of both histological types together, peaking at age 35, with relative risk at 156.3 (Weibull) and 214.8 (NP) and decreasing to 25.4 and 9.5 at age 65.

It is possible to observe a decrease in both cumulative and relative ratio risk estimations when leaving pT1a lesions out of DGC event consideration.

The observed differences between both models – for either pT1a consideration or not – reveals that choice of model will probably not influence the outcome of the risk estimations by much, at least until after 50, be it cumulative or relative risk.

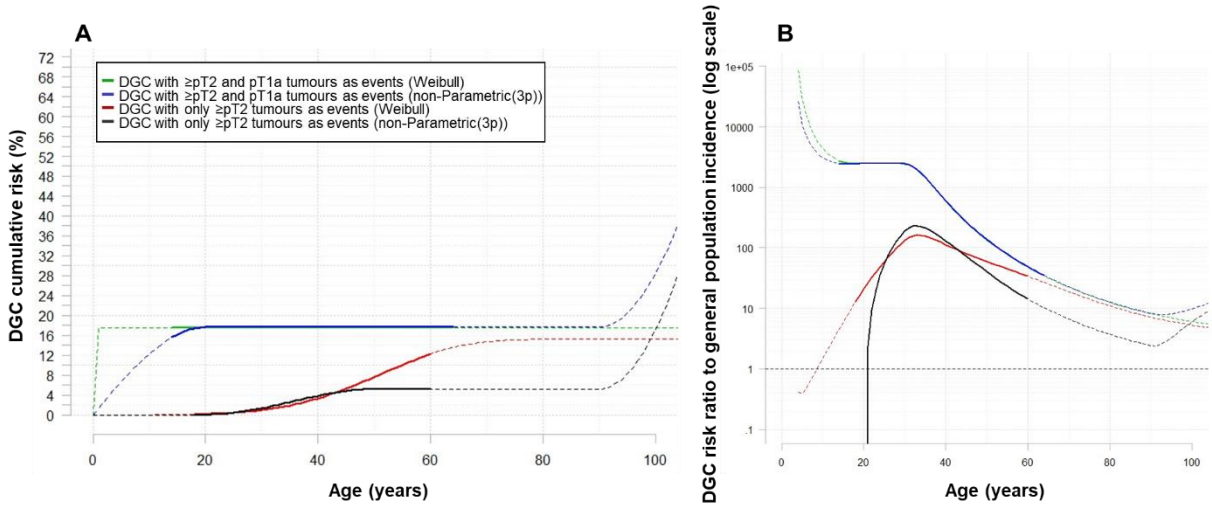


Figure 12. GRL risk estimations for DGC onset. A. Cumulative risk for DGC onset for both pT1a and ≥pT2 tumours (Weibull – green; NP – blue). Cumulative risk for DGC onset for ≥pT2 tumours only (Weibull – red; NP – black). **B.** Risk ratio for DGC onset for both pT1a and ≥pT2 tumours in HDGC against general population (Weibull – green; NP – blue). Risk ratio for DGC onset for ≥pT2 tumours only in HDGC against general population (Weibull – red; NP – black).

Slight differences in the risk curves for the same condition may be the result of the way both models behave. The Weibull function, as a parametric model, assumes that the model shape will behave in specific ways. This makes it less robust in cases with a small sample size or with some instability in the event distribution within the cohort.

Lobular Breast Cancer (LBC)

Regarding the LBC risk estimations, the methods applied demonstrate very close estimations in both the cumulative risk and risk ratios to population (Figure 13, panels A & B).

Cumulative risk with the Weibull model follows an almost constant pattern, reaching a cumulative risk estimation of 11.8% at age 65. NP predicts very similar numbers as the Weibull model, fluctuating between higher and lower values in comparison across the different ages, and estimates a risk of 11.5% at age 65. After 61 years, no more cases were reported in the cohort, but it is possible to observe that the Weibull model predicts a continuous increase, while NP model stagnates around age of 80 years.

HDGC-related LBC relative risk for these families followed almost the same curve in both both Weibull and the 3-points NP models as well. The relative risk to general population at 30 years,

where it peaks was 243.1 and 212.2 with Weibull and 3-points NP, respectively. Both curves show a decrease in relative risk, reaching around 27 in both models' estimations.

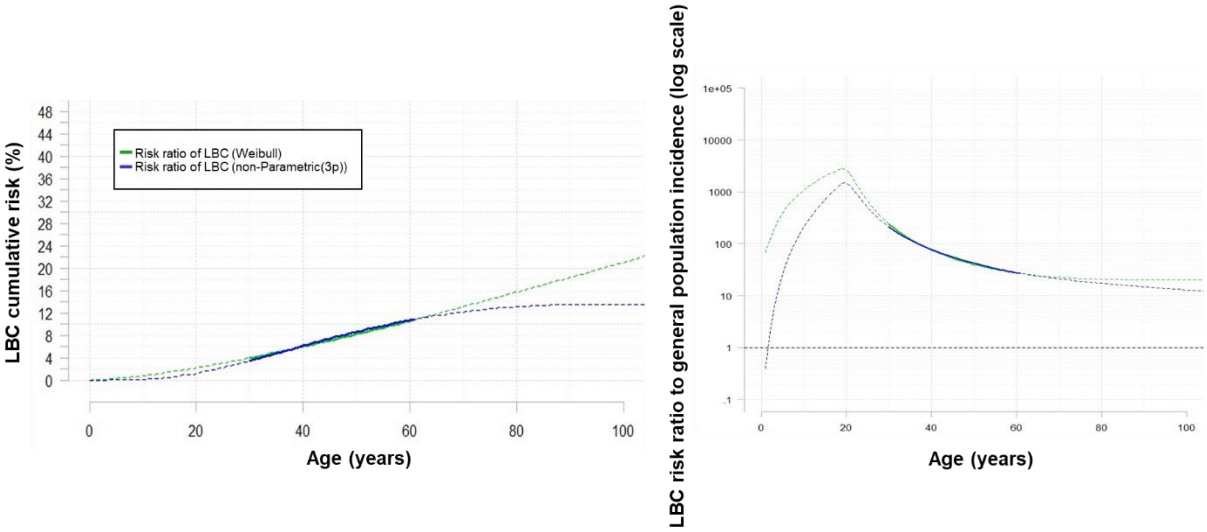


Figure 13. GRL risk estimations for LBC onset. A. Cumulative risk for LBC onset (Weibull – green; NP – blue). **B.** Risk ratio for LBC onset in HDGC against general population (Weibull – green; NP – blue).

Sensitive Analysis

A sensitive analysis was performed to verify if one of the families was an outlier in the risk analyses. Risk curves were generated, each without one of the families. Both Weibull and 3-points NP models were evaluated, with DGC cases comprising only ≥pT2 tumours.

In the Weibull model analysis both cumulative and relative risk display similar curves when considering out each of the families. In the cumulative risk estimations, the most notorious differences manifest after 40 years, with F1 out diminishing the risk considerably, and F2 and F5 out returning curves with a higher risk ratio (Figure 14, panel A). In the relative ratio to Northern Portuguese population, leaving F1, F4 and F11 out analyses are the most different from the trend, starting with a higher risk at younger ages and then declining rapidly from 30 years onwards(Figure 14, panel B). The weight of these families can be associated with the extent of genetic testing and patients kept under surveillance of each family.

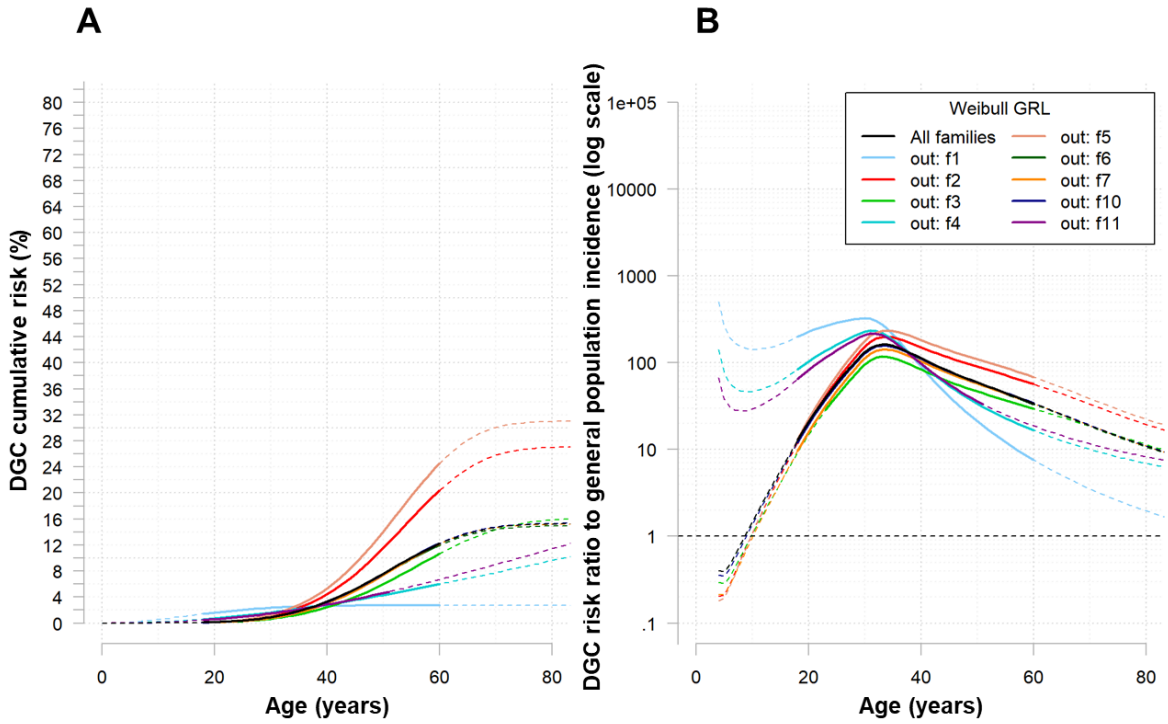


Figure 14. Leave-one-out sensitive analysis for risk estimations of DGC onset. A. Cumulative risk for DGC onset considering only ≥pT2 tumours in black. A curve for every estimation leaving one of the families out of consideration. **B.** Risk ratio for DGC onset considering only ≥pT2 tumours in HDGC against general population in black. A curve for every estimation leaving one of the families out of consideration.

The non-parametric method showed evenly curves in the risk ratio analysis, with one family left out being very different from the rest (Supplementary Figure 1). Results are attenuated compared with the Weibull model, which is attributed to the more flexibility and less distribution assumptions.

LBC leave-one-out sensitive analysis was not performed due to low sample size.

4.3. RNA-sequencing – A valuable output from the developed HDGC clinical database

Histological samples from tumoral, metastatic, and normal tissue of DGC patients followed and treated at CHUSJ was used for RNA extraction and subsequent RNA-sequencing analysis.

This analysis was performed with 30 samples from 16 patients. These were distributed among 11 normal tissue samples, 9 pT1a samples, 4 ≥pT2 tumour samples, and 6 tissue from metastasis (5 long-distance, 1 from lymph node). It was possible to obtain paired samples (tumour + normal) from 9 individuals with pT1a foci and 2 with ≥pT2 tumour.

All the individuals were known carriers of the *CDH1* c.1901C>T variant from the HDGC cohort.

RNA-sequencing analysis was performed using e-Bayes ANOVA algorithm, and log2 Fold-Change threshold was established as below -2 or above 2. A p-value of <0.01 for significant differentially expressed genes was considered. No p-value adjustment was introduced in the analysis between pair of tissue types, due to the low sample size and no revelatory output.

RNA-sequencing for LBC was not performed due to the low number of samples at disposal until to this date. More RNA samples are being extracted, both from LBC patients' breast tissue and from carriers that performed risk-reducing mastectomy without lesions (Figure 7).

4.3.1. Metastatic and advanced tumour tissue samples group well defined clusters; normal and pT1a samples mix amongst themselves

In a first step, a PCA mapping was performed to check for possible clusters of the studied samples. The four metastasis and four ≥pT2 separated into two same tissue groups, in proximity, but well established as two different groups (Figure 15).

Findings in the PCA analysis reveal that there is no distinction in the gene expression between pT1a and normal tissue. Furthermore, some samples are scattered across the PCA plot, with six (3 normal gastric tissue, 3 pT1a) being well apart from the rest. These samples belong to the same three patients, however there was no other known batch effect that could be exclusive of these (RNA concentration, type, and time of extraction).

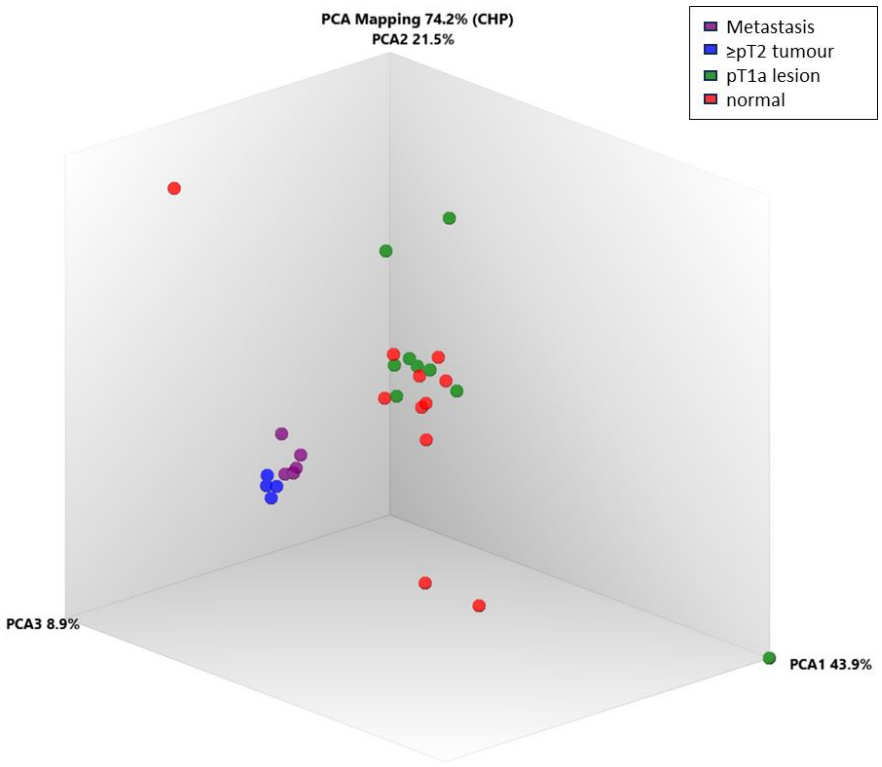


Figure 15. Principal Component Analysis of 30 RNA-sequencing samples from 16 HDGC patients

Further analyses corroborated the PCA findings. An heatmap with all 30 samples and an F-Test value of <0.01 revealed 586 genes in significant variance between the four conditions of the analysis (Figure 16). Additionally, there was a clear distinction between two main groups. Metastasis and ≥pT2 samples grouped in one part of the heatmap (each tissue type in its own subgroup), while pT1a lesions and normal gastric tissue clustered together and among themselves.

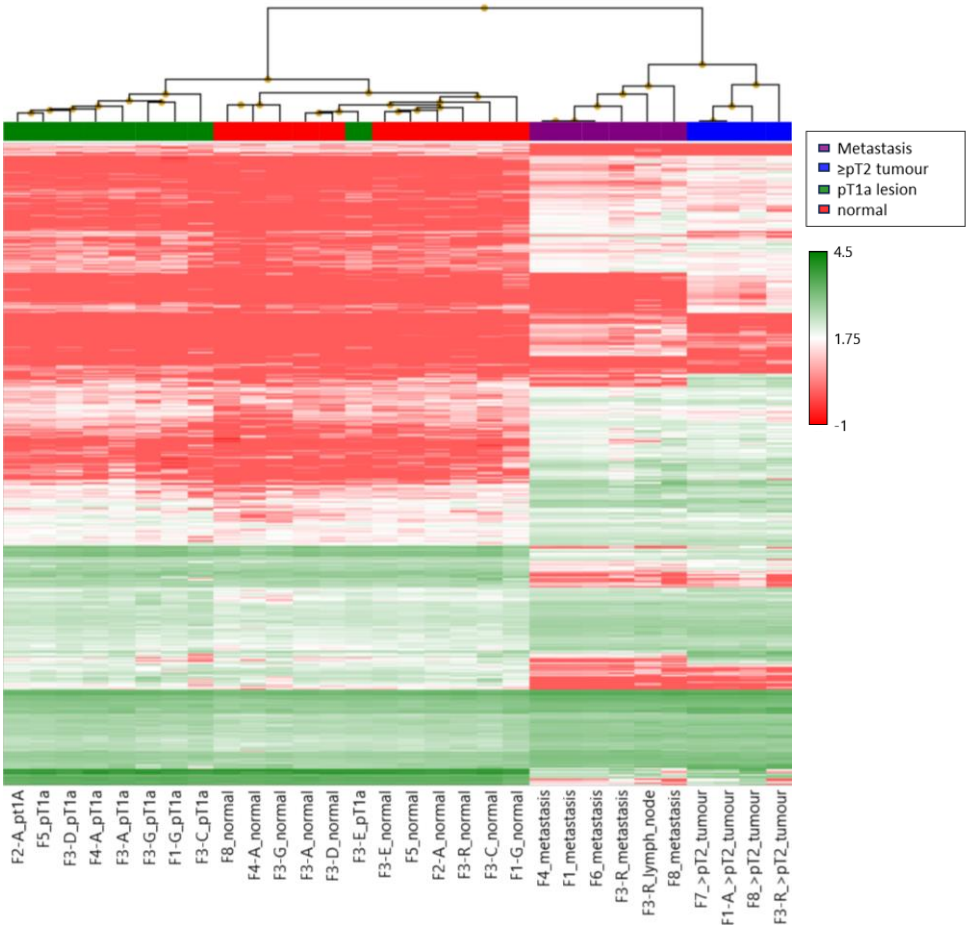


Figure 16. Heatmap of 30 samples from 16 HDGC patients comparing four different tissue classifications. Log2 transformed scale of gene expression per sample. Colour gradient red-to-green from low to high expression values.

Selecting genes with an FDR F-test below 0.05, there was a 20-gene output with some interesting findings. It is possible to observe the downregulation of *MUC6* and *PSCA* in the ≥pT2 and metastatic tissue, while *UCA1*, a lncRNA associated with cancer progression and metastasis, is slightly upregulated in the same tissue (Table 8).

Table 8. Differentially expressed genes in a four-way comparison of the gastric tissue types and metastasis

ID	Metastasis Avg (log2)	≥pT2 Avg (log2)	pT1a Avg (log2)	Normal Avg (log2)	F-Test	FDR F-Test
EPN3	2.58	3.54	6.64	6.34	1.19E-07	0.0015
CKB	5.63	8.54	10.43	10.37	1.40E-07	0.0015
PDIA2	1.67	2.22	7.12	7.34	5.81E-07	0.0033
MUC6	2.37	5.17	10.28	9.77	6.41E-07	0.0033
FLJ42875	1.18	1.68	6.18	6.65	2.28E-06	0.0082
CHRNA3	-0.03	2.86	0.16	0.12	2.62E-06	0.0082
CLIC6	3.21	4.64	8.57	8.47	2.77E-06	0.0082
UCA1	3.06	2.44	-0.06	-0.12	4.04E-06	0.0100
LGALS9B	1.2	1.83	5.92	5.79	4.32E-06	0.0100
RAP1GAP	3.62	4.13	7.96	8.03	5.09E-06	0.0100
IGF2BP3	3.63	2.9	0.21	0.17	5.30E-06	0.0100
CPA2	0.26	1.29	7.7	7.5	6.49E-06	0.0112
SLC5A5	0.67	1.95	6.7	6.39	7.09E-06	0.0113
MFSD4	1.31	3.43	7.41	7.33	9.71E-06	0.0136
DPCR1	1.38	4.63	8.93	8.61	9.82E-06	0.0136
PSCA	2.81	4.64	10.84	11.06	1.31E-05	0.0170
MARCH3	3.96	4.01	0.82	1.01	1.44E-05	0.0176
KCNQ3	2.26	2.8	-0.22	-0.22	1.62E-05	0.0187
SLC26A9	2.05	2.57	6.06	5.8	3.67E-05	0.0402
TACR2	0.14	6.77	1.43	1.6	4.55E-05	0.0474

These findings are all preliminary, the sample size is small compared with a gene panel of nearly 21,000 genes, and RNA is still being collected from FFPE blocks at the time of this writing.

However, these first analysis revealed some interesting takes for the future, and it was decided to analyse which pathways may be enriched in comparisons between each pair of tissue type.

4.3.2. DGC enriched pathways

Pathway enrichment analysis was performed with a selection of genes differentially expressed between each pair of tissue type, with a p-value threshold of <0.01. Enriched pathways were retrieved with a <0.1 threshold of FDR correction.

Enriched pathways depicted in Figure 17 were a selection from the filters mentioned above, to highlight important pathways for DGC tumour environment, proliferation, and invasion. It is possible to observe dysregulation of pathways such as cell-matrix adhesion, focal adhesion, ERK1/ERK2 cascade, and serine/threonine kinase signalling pathway. Most of these pathways appear dysregulated in the ≥pT2 vs normal gastric tissue comparison.

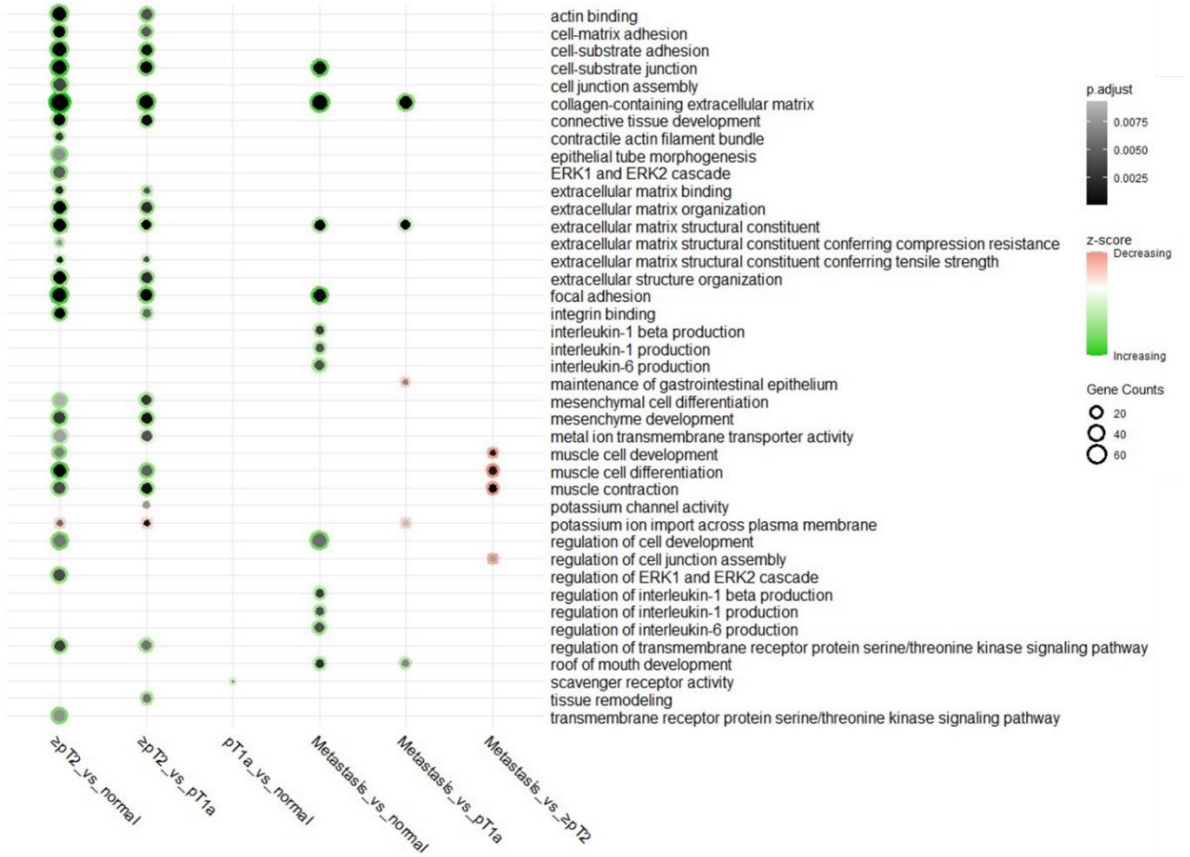


Figure 17. Selected enriched pathways for each sample type comparison. Enriched pathways with FDR p-value <0.1. Size of FDR p-value inversely proportional to its true value. Colour gradient red-to-green for z-score (decrease to increase).

5. Discussion

Hereditary cancer research is a multifaceted process that requires the full cooperation of multidisciplinary teams to have fruition. It is worth noticing that a valuable and robust output is only possible if in every step of the process, data can be safeguarded, obtained, and transmitted properly. In the presented work, it is possible to observe the power that missing data can have on the final analysis, the limitations of the available resources, and all the needed care to handle sensitive data and how to process it to present results in full anonymity.

As stated in the results section, this work highlights the importance of well collected and curated data for scientific research and health care providing. Data from 11 families was accessed – that later became 9 – that is being collected since the first isolated case of early onset DGC that was carrying the c.1901C>T variant was identified in Northern Portugal, in 2003[14]. Data that was collected across a 20-year span, by different health professionals, with the multiple new guidelines and updates that were introduced either in criteria for genetic testing, management, and surveillance of HDGC, or histopathological records' specifications[30,56,74,88]. Some data was also collected in retrospective from before 2003, as more families were ascertained, and past cases of gastric cancer were identified within these families. Therefore, it is important to be aware of the limitation of data availability for some of the individuals and the incomplete information of certain variables assessed in the compiled database. Nevertheless, in a context of an hereditary rare disease and comparing to other available datasets this work established the largest dataset for a specific variant that is the most recurrent variant in Europe[63,89].

At the start, the information of 400 individuals was gathered, from which it was apparent that sufficient information would not be possible to obtain to include some of these in the analysis pipeline. The most obvious gap were the individuals identified through patients' hearsay. The database was therefore trimmed to 285, to include individuals with at least information of age of last news for DGC and/or LBC, the minimum information required for penetrance and risk estimation analysis. CHUSJ was a source of vast amount of more detailed data, including confirmed genetic tests for the c.1901C>T variant, pathways of surveillance and management of carriers within the institution, histopathological reports from biopsies, gastrectomy, and mastectomy tissue.

Although the *CDH1* c.1901C>T variant is an ultrarare variant with no population frequency information in clinical variants databases[90], the observed population frequency of the variant in the CHUSJ cohort from North of Portugal is 43.8% and 33.0% in the tested and extended database individuals, respectively. Certainly, it has to be taken into consideration that this is a familial study and focused on a singular condition which, in conjunction with previous studies identifying this variant as pathogenic/likely-pathogenic and the considerable frequency of DGC and LBC cases among these families, highlights the functional impact this variant has on the development of HDGC[52,54,63,64,91].

However, as it is possible to observe in this familial cohort, the penetrance of disease is clearly incomplete, with only 13.2% of positive carriers reported at CHUSJ (or 48% if one considers pT1a lesions found in RRG) displaying DGC, and 15% of female carriers developing LBC. This suggests some possible external factor to disease onset, either it being environmental or, as suggested in literature, possible co-segregation of genetic modifiers, that may be either protecting against or promoting disease onset[57,58,63]. It also reveals that better estimates could be achieved if performed without the one-size-fits-all approach form previous studies. It should be taken into account the peculiarities of each family. The family history, which variant is present, how it segregates, the incidence of cancer within the family and its age of onset.

Interesting findings in the cohort were the presence of superficial and chronic gastritis, and *Helicobacter pylori* across all individuals. Although statistically significant correlation was not found between any of these findings and cancer onset (or not), it was observed that more than half of the individuals with pT1a lesions at biopsy or at RRG presented, at some surveillance intervention, both conditions. Additionally, the same applies to carriers undergoing surveillance but showed no lesions and chose to not perform RRC until this date. Although most of the studies find no significant association of *H. pylori* infection and HDGC[5,36,39], it is still a subject under study in family gastric cancer onset[43,92–94], and eradication is advisable in CDH1 variant carriers[30].

Fisher's exact test applied to the clinical findings of this cohort suggest no association is evident between *H. pylori* (and gastritis) and DGC development. Although the number of infection cases is high among pT1a cases, it is also present in approximately 55% of non-affected individuals. This seems to be in line with the reported high incidence of *H. pylori* in the Portuguese population[95]. However, cautious conclusions should be taken from these analyses. The cohort sample size for a robust statistical test was small and many patients had incomplete health records registry or histopathological reports of gastric biopsy/RRG tissue. Most of the advanced cancer patients had no information about underlying conditions (such as gastritis, *H. pylori* infection, intestinal metaplasia, or other assorted findings). Some of this stems from the fact that reports were written in paper and thus lost to digital survey (only the cancer stage was introduced) and other because only the main finding was reported. The same happened in some biopsy and RRG reports from pT1a patients.

It is important to tackle the way that the histopathology reports – and, as an extension, the clinical health records – are compiled for better management of useful information, either to clinic or research. There is no question that the most important function in healthcare is the life and quality of life of the patient. Direct care to the patient should and always will be the main priority. However, small adjustments to the indirect care could be performed to ensure that the patient, families, and general population could benefit in long term. Introduction of defined tick box fillings in reports for pertinent secondary findings could be a solution for some data. Even if there is no possible confirmation or not of certain condition, a "not applicable/not available" option is more informative than none in retrospective studies that may be using that data later.

Surveillance context of LBC in the CHUSJ group follows the criteria established in the most recent guidelines for HDGC[30], with female carriers performing MRI in annual intervals, and adding mammography and ultrasound from the age of 40 years[30]. Biopsies are only performed after a suspicion mass is detected, so little histological information is obtainable from reports. From 10

women with biopsy reports, seven have performed mastectomy (five of them bilateral), with four cases of pT1c or higher, one of unknown histology and 3 LIN findings, with only one perishing from consequences of LBC. Only three women without any LBC warning sign decided to undergo RRM. This reveals that breast monitoring has been quite successful in early detection and carefully following any suspicion of malignancy among HDGC families at CHUSJ. However, there was a case of misclassification, the proband of F9 (later annexed to F2), with *in situ* lesion detected in one of the breasts and finding with unknown malignancy potential on the other during biopsy, was diagnosed with LBC and submitted to bilateral mastectomy. It was later reported both breasts had LIN findings, after mastectomy tissue assessment. The procedure was well evaluated, and the mastectomy would be the right decision, in context of surveillance, considering the woman showed premalignant lesions and was 44 years old, where mastectomy is not unadvisable. The woman was the first case of that family branch to be identified as *CDH1* c.1901C>T variant carrier, and thus the decision to perform bilateral mastectomy was probably correct, besides the LBC misdiagnosis.

Cumulative and lifetime risk estimations for DGC and LBC for carriers of *CDH1* the c.1901C>T variant in the context of HDGC have been calculated in several publications and different family populations for the past 20 years[54,57–59]. These have been evolving with time, becoming less biased through family ascertaining and cohort composition, and both the obtained risk and range of 95% confidence intervals have been steadily decreasing from 67% (95%CI, 39%-99%) cumulative risk of DGC at age 80 for males and 83% (95%CI, 58%-100%) for women, described in Pharoah *et al* in 2001[57], to 42% (95%CI, 30%-56%) in men and 33% (95%CI, 21%-43%) in women, in 2019[59]. LBC cumulative risk for women increased, although it became more precise, going from 39% (95%CI, 12%-84%) in 2001, to 55% (95%CI, 39%-68%) in 2019. It is important to emphasize that all risk estimations have their own cohort of study, calculation methods and limitations. One that is transversal to all, and probably being mitigated with time is the bias towards people that are tested for *CDH1* variants.

Another possible limitation of the reported penetrance and risk studies for HDGC is the use of families sharing different *CDH1* pathogenic variants[54,57–59]. This work represents a big shift from previous studies as it is the first that has the goal to develop lifetime risk estimations for a single *CDH1* variant – c.1901C>T – shared between nine different families, all from the same region in North of Portugal. As a region with a well-established industrial sector and economic stability within the country, families tend to settle and rarely leave, which may have been the reason to find so many families sharing the same variant, establishing a founder effect[63,65].

First risk analysis was performed with Kaplan-Meier non-parametric statistical method. There were three cumulative risk analyses developed, one for DGC considering both ≥pT2 and pT1a lesions as cancer events, while another one only considering ≥pT2 as an event. The third estimation was performed for LBC in women from the cohort. Kaplan-Meier risk estimates revealed a cumulative risk at age 65 of approximately 60% and 19% for c.1901C>T carriers considering both ≥pT2 and pT1a as events or just ≥pT2, respectively. Individuals with unknown genotype had approximately 27% or cumulative risk at age 65. Log-rank tests reveal a statistically significative higher risk when accounting with pT1a lesions as cancer events together with advanced tumour,

compared to ≥pT2 alone. For LBC, cumulative risk of known carriers of the c.1901C>T variant was around 34% at 65.

These results were corroborated with the penetrance estimations performed with the Genotype Restricted Likelihood (GRL) method developed by Bonaïti *et al* in 2011[77]. It was already used and with output publications of risk estimates for variants in *MLH1*, *MSH2* and *MSH6* genes in the context of Lynch syndrome[96], and *CTNNA1* variant carriers risk estimations for DGC[76].

Risk estimations with GRL have revealed a cumulative risk of around 13.8% (Weibull model) at age 65 for cancer being considered only advanced tumour staging, and a risk of 17.5% when both pT1a findings and ≥pT2 tumour are considered cancer. Both models maintain a similar increase of risk across age until 45 years old, where NP model stays constant and Weibull model continues to increase based on function assumptions.

DGC risk ratios against the Northern Portuguese population reveal a peak of relative risk between 20 and 30 years considering pT1a and ≥pT2 tumours as DGC, with both models behaving similarly (Figure 12, panel B). Risk estimations at age 30 are 2445.5 and 2466.7, decreasing to 32.3 and 32.5 at age 65, from Weibull and non-parametric models respectively. When considering only ≥pT2 tumours as cancer, the relative risk to population is much lower than the calculation of both histological types together, peaking at age 35, with relative risk at 156.3 (Weibull) and 214.8 (NP) and decreasing to 25.4 and 9.5 at age 65. Cumulative risk for LBC in women retrieved was 11.8% at 65 years old.

Estimations from both Kaplan-Meier curves and through the GRL method reveal that in the North of Portugal cohort have a cumulative risk lower than what is reported in literature. It is possible that these estimations are lower for a few different reasons. This study is the first one where risk and penetrance estimates are being calculated for a single *CDH1* variant, following families from the same geographical region, with a total of 11 starting probands, 171 tested individuals, known genotype of 29 more, and 85 individuals considered from hearsay, older CHUSJ reports, or confirmation by other institutions. Some of these families were already being followed for decades at the institution, with tests performed across generations, and without bias towards affected patients. Of course, there is always the bias that all these tests occur within the context of HDGC, the population of study will always be within the boundaries of hereditary cancer. However, that was the goal that was behind this thesis and, as an extent, the FCT-funded project where it stems from; to estimate intrafamilial lifetime risk for *CDH1* carriers in the context of HDGC. With risk estimations of 19.0% in Kaplan-Meier calculations and 13.8% with the GRL Weibull method, there is a strong incomplete penetrance of disease in carriers of the c.1901C>T variant across these families. An outcome that could be attributed to genetic modifiers, that may be influencing the development of disease either with a protective or enhancing effect.

Overall, cumulative risk ratios obtained in this study, suggest that the *CDH1* c.1901C>T variant has a lower penetrance and risk for HDGC disease development than what has been reported in other studies for *CDH1* variants.

Risk ratio calculations with GRL and Standard Incidence Ratio (SIR), also reveal that the probability of developing DGC at a younger age of onset is much higher in HDGC families when compared to the general population. Specially between 20 years and 35 the relative risk of developing HDGC is 156 higher than in population from Northen Portugal by the GRL method

and 34 times higher considering SIR calculations until 35 years. These findings are concurrent with what is attributed to DGC in the context of HDGC, this disease strikes at an early age of onset[30,53].

Furthermore, across this work, estimations with pT1a lesions as cancer events were also considered. It is possible to observe an increase in the risk estimations, with numbers such as 60% (Kaplan-Meier) and 17.5 (GRL Weibull) of risk at 65 years. It was one of the goals to observe how much these lesions would influence risk estimations for DGC and to understand their role and possible meaning withing the context of HDGC surveillance. It is important to notice that, however higher the risk considering pT1a lesions, these are indolent lesions with uncertainty of disease progression, and surveillance procedures when in the light of these findings is still a hot topic[64,97].

It is established that pT1a findings at endoscopy with random biopsies patients from HDGC families under surveillance, are advised to remove the stomach as a preventive measure[3,30,39,97]. However, risk reducing gastrectomy (RRG) is an invasive method with possible complications and morbidity consequences[98,99]. It is therefore important to assess the risk of DGC development in the presence of pT1a lesions, and weight the better solutions for each particular case. Van der Post *et al* proposes a possible modification of surveillance protocol, advising on a more thorough investigation on the type of pT1 lesion found, with focus on lesions that would suggest expansion towards the submucosa layer[97].

In the RNA-sequencing preliminary findings, it is possible to observe that, when comparing 30 samples from these c.1901C>T variant sharing families, pT1a samples tend to group together with normal tissue samples in both PCA and hierarchical clustering outputs. In fact, applying more strict filters to the differential expression (DE) analysis (considering FDR p-value adjustment of <0.1), no genes are found to be differentially expressed between pT1a and normal tissue. Taking the genes from Table 8 from the results section, it is possible to observe that log2 of average signal in normal and pT1a samples is constantly similar among the genes with highest statistical DE in the comparison across all 4 tissue types. There is the possibility of many of these pT1a lesions used for RNA-sequencing analysis being of small size, and with normal tissue contamination, but it is interesting to find such high similarity between samples of both tissues.

The goal of this work is not to suggest that pT1a lesions should not be considered a risk factor for DGC development, but, in line with Van der Post *et al*[97], that a more critical interpretation of these lesions should be performed before the enactment of more invasive preventive measures. RRG is still the recommended procedure for *CDH1* pathological variant carriers with history of DGC within the family[30]. To notice, most of the gastrectomy procedures found in the families of the present study are RRG without previous findings in gastric biopsy. It is important to reflect on the power that findings of pT1a lesions have in determining which preventive path to follow for each patient, and if, and how, could a better stratification based on histology of these lesions be beneficial for patient care.

Preliminary RNA-sequencing analysis reveals some interesting pathway enrichment between advanced tumour and normal tissue samples. It is possible to observe that many pathways related to cell-cell adhesion, extracellular matrix organization, the Mitogen-Activated Protein Kinase/Extracellular Signal-Regulated Kinase (MAPK/ERK) pathway are being disrupted when

comparing advanced tumour to normal (and in some, pT1a) tissue. Further exploration of this data is underway, with the intent to also add more samples, but it is already interesting to find that many pathways related to processes such as cell growth, differentiation, and immune response are being statistically significantly disrupted in tumour tissue.

6. Conclusion

This work highlights the importance of a well-established and robust database of clinical and genetic data in the study of HDGC, with the goal to improve patient care and surveillance methods. Three different output analysis were possible with information from the established database, and further outcomes are underway.

It was possible to characterise the CHUSJ families with the *CDH1* c.1901C>T variant, in terms of distribution of carriers, disease events, groups of interest, histopathological findings, and surveillance procedures, while developing extended pedigrees for each of the families.

Lifetime risk estimations with the GRL method revealed that the cumulative risk of DGC in the c.1901C>T cohort was 13.8 at age 65, while LBC had a cumulative risk of 11.5 at the same age. These findings suggest a lower risk for carriers of this specific *CDH1* variant compared to previous reports from lifetime risk estimations of several *CDH1* variants. This was the first intrafamilial risk study consisting of one single variant across families from the same region.

A preliminary RNA-sequencing analysis was performed, revealing almost no differences between expression patterns of pT1a and normal tissues, opening the door to explore the significance of these lesions in surveillance of HDGC. Pathways involved in cell growth, differentiation and immune response were also found to be dysregulated in advanced tumour tissue (≥pT2), instigating further studies on the transcriptomic landscape of HDGC.

7. Future Work

Further studies to RNA-sequencing data are underway, where it will be explored the nature of the similarities between pT1a and normal tissue found on the preliminary analysis. It will also be assessed the significance of the enriched pathways in the advanced tumour tissue, specially how it possibly associates to cancer growth, invasiveness, and aggressiveness.

Genomic data is also being explored, where studies are being performed to try to reveal genetic modifiers in selected groups of patients that could be having a protective or enhancing effect on the outcome of disease progression in carriers of the c.1901C>T variant. Groups of interest were established bases on age and onset (or not) of the disease.

This database with serve as the basis for all studies relevant for HDGC analysis in these families, be it either statistical, surveillance or treatment studies, so the continuous imputation of new data from these families, within ethical committee permissions, will be a priority.

8. References

1. Thrift AP, Wenker TN, El-Serag HB. Global burden of gastric cancer: epidemiological trends, risk factors, screening and prevention. *Nat Rev Clin Oncol*. 2023;20(5):338-349. doi:10.1038/s41571-023-00747-0
2. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 2021;71(3):209-249. doi:10.3322/caac.21660
3. Pilonis ND, Tischkowitz M, Fitzgerald RC, di Pietro M. Hereditary Diffuse Gastric Cancer: Approaches to Screening, Surveillance, and Treatment. *Annu Rev Med*. 2021;72(1):263-280. doi:10.1146/annurev-med-051019-103216
4. Iyer P, Moslim M, Farma JM, Denlinger CS. Diffuse gastric cancer: histologic, molecular, and genetic basis of disease. *Transl Gastroenterol Hepatol*. 2020;5:52-52. doi:10.21037/tgh.2020.01.02
5. Oliveira C, Pinheiro H, Figueiredo J, Seruca R, Carneiro F. Familial gastric cancer: genetic susceptibility, pathology, and implications for management. *Lancet Oncol*. 2015;16(2):e60-e70. doi:10.1016/S1470-2045(14)71016-2
6. Lauwers GY, Mullen JT, Chelcun Schreiber KE, Chung DC. Familial Gastric Cancers. *Pathol Case Rev*. 2014;19(2):66-73. doi:10.1097/PCR.0000000000000030
7. Machlowska J, Baj J, Sitarz M, Maciejewski R, Sitarz R. Gastric Cancer: Epidemiology, Risk Factors, Classification, Genomic Characteristics and Treatment Strategies. *Int J Mol Sci*. 2020;21(11):4012. doi:10.3390/ijms21114012
8. Decourtye-Espiard L, Guilford P. Hereditary Diffuse Gastric Cancer. *Gastroenterology*. 2023;164(5):719-735. doi:10.1053/j.gastro.2023.01.038
9. Gamble LA, Heller T, Davis JL. Hereditary Diffuse Gastric Cancer Syndrome and the Role of *CDH1*. *JAMA Surg*. 2021;156(4):387. doi:10.1001/jamasurg.2020.6155
10. Oliveira C, Pinheiro H, Figueiredo J, Seruca R, Carneiro F. E-Cadherin Alterations in Hereditary Disorders with Emphasis on Hereditary Diffuse Gastric Cancer. In: *Progress in Nucleic Acid Research and Molecular Biology*. ; 2013:337-359. doi:10.1016/B978-0-12-394311-8.00015-7
11. Frebourg T, Oliveira C, Hochain P, et al. Cleft lip/palate and CDH1/E-cadherin mutations in families with hereditary diffuse gastric cancer. *J Med Genet*. 2005;43(2):138-142. doi:10.1136/jmg.2005.031385
12. Milne AN, Offerhaus GJA. Early-onset gastric cancer: Learning lessons from the young. *World J Gastrointest Oncol*. 2010;2(2):59. doi:10.4251/wjgo.v2.i2.59
13. Guilford P, Hopkins J, Harraway J, et al. E-cadherin germline mutations in familial gastric cancer. *Nature*. 1998;392(6674):402-405. doi:10.1038/32918
14. Suriano G, Oliveira C, Ferreira P, et al. Identification of CDH1 germline missense mutations associated with functional inactivation of the E-cadherin protein in young gastric cancer probands. *Hum Mol Genet*. 2003;12(5):575-582. doi:10.1093/hmg/ddg048
15. Kaurah P, Huntsman DG. *Hereditary Diffuse Gastric Cancer*.; 1993.

16.Lynch HT, Grady W, Suriano G, Huntsman D. Gastric cancer: New genetic developments. *J Surg Oncol*. 2005;90(3):114-133. doi:10.1002/jso.20214
17.Bacani JT, Soares M, Zwingerman R, et al. CDH1/E-cadherin germline mutations in early-onset gastric cancer. *J Med Genet*. 2006;43(11):867-872. doi:10.1136/jmg.2006.043133
18.van Roy F, Berx G. The cell-cell adhesion molecule E-cadherin. *Cellular and Molecular Life Sciences*. 2008;65(23):3756-3788. doi:10.1007/s00018-008-8281-1
19.Oliveira C, de Bruin J, Nabais S, et al. Intragenic deletion of CDH1 as the inactivating mechanism of the wild-type allele in an HDGC tumour. *Oncogene*. 2004;23(12):2236-2240. doi:10.1038/sj.onc.1207335
20.Shenoy S. CDH1 (E-Cadherin) Mutation and Gastric Cancer: Genetics, Molecular Mechanisms and Guidelines for Management . *Cancer Manag Res*. 2019;Volume 11:10477-10486. doi:10.2147/CMAR.S208818
21.Wong SHM, Fang CM, Chuah LH, Leong CO, Ngai SC. E-cadherin: Its dysregulation in carcinogenesis and clinical implications. *Crit Rev Oncol Hematol*. 2018;121:11-22. doi:10.1016/j.critrevonc.2017.11.010
22.Valkenburg KC, Graveel CR, Zylstra-Diegel CR, Zhong Z, Williams BO. Wnt/β-catenin Signaling in Normal and Cancer Stem Cells. *Cancers (Basel)*. 2011;3(2):2050-2079. doi:10.3390/cancers3022050
23.Hong Y, Manoharan I, Suryawanshi A, et al. β-Catenin Promotes Regulatory T-cell Responses in Tumors by Inducing Vitamin A Metabolism in Dendritic Cells. *Cancer Res*. 2015;75(4):656-665. doi:10.1158/0008-5472.CAN-14-2377
24.Jeanes A, Gottardi CJ, Yap AS. Cadherins and cancer: how does cadherin dysfunction promote tumor progression? *Oncogene*. 2008;27(55):6920-6929. doi:10.1038/onc.2008.343
25.Mendonsa AM, Na TY, Gumbiner BM. E-cadherin in contact inhibition and cancer. *Oncogene*. 2018;37(35):4769-4780. doi:10.1038/s41388-018-0304-2
26.Mateus AR, Simões-Correia J, Figueiredo J, et al. E-cadherin mutations and cell motility: A genotype–phenotype correlation. *Exp Cell Res*. 2009;315(8):1393-1402. doi:10.1016/j.yexcr.2009.02.020
27.Sommer AK, te Paske IBAW, Garcia-Pelaez J, et al. Solving the genetic aetiology of hereditary gastrointestinal tumour syndromes– a collaborative multicentre endeavour within the project Solve-RD. *Eur J Med Genet*. 2022;65(5):104475. doi:10.1016/j.ejmg.2022.104475
28.Garcia-Pelaez J, Barbosa-Matos R, São José C, et al. Gastric cancer genetic predisposition and clinical presentations: Established heritable causes and potential candidate genes. *Eur J Med Genet*. 2022;65(1):104401. doi:10.1016/j.ejmg.2021.104401
29.Weren RDA, van der Post RS, Vogelaar IP, et al. Role of germline aberrations affecting *CTNNA1* , *MAP3K6* and *MYD88* in gastric cancer susceptibility. *J Med Genet*. 2018;55(10):669-674. doi:10.1136/jmedgenet-2017-104962
30.Blair VR, McLeod M, Carneiro F, et al. Hereditary diffuse gastric cancer: updated clinical practice guidelines. *Lancet Oncol*. 2020;21(8):e386-e397. doi:10.1016/S1470-2045(20)30219-9

31. Aronson M, Swallow C, Govindarajan A, et al. Germline Variants and Phenotypic Spectrum in a Canadian Cohort of Individuals with Diffuse Gastric Cancer. *Current Oncology*. 2020;27(2):182-190. doi:10.3747/co.27.5663
32. LAURÉN P. THE TWO HISTOLOGICAL MAIN TYPES OF GASTRIC CARCINOMA: DIFFUSE AND SO-CALLED INTESTINAL-TYPE CARCINOMA. *Acta Pathologica Microbiologica Scandinavica*. 1965;64(1):31-49. doi:10.1111/apm.1965.64.1.31
33. Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA Cancer J Clin*. 2017;67(2):93-99. doi:10.3322/caac.21388
34. Huntsman DG, Carneiro F, Lewis FR, et al. Early Gastric Cancer in Young, Asymptomatic Carriers of Germ-Line E-Cadherin Mutations. *New England Journal of Medicine*. 2001;344(25):1904-1909. doi:10.1056/NEJM200106213442504
35. Monster JL, Kemp LJS, Gloerich M, van der Post RS. Diffuse gastric cancer: Emerging mechanisms of tumor initiation and progression. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*. 2022;1877(3):188719. doi:10.1016/j.bbcan.2022.188719
36. Carneiro F, Huntsman DG, Smyrk TC, et al. Model of the early development of diffuse gastric cancer in E-cadherin mutation carriers and its implications for patient screening. *J Pathol*. 2004;203(2):681-687. doi:10.1002/path.1564
37. Iyer P, Moslim M, Farma JM, Denlinger CS. Diffuse gastric cancer: histologic, molecular, and genetic basis of disease. *Transl Gastroenterol Hepatol*. 2020;5:52-52. doi:10.21037/tgh.2020.01.02
38. Tsugeno Y, Nakano K, Nakajima T, et al. Histopathologic Analysis of Signet-ring Cell Carcinoma In Situ in Patients With Hereditary Diffuse Gastric Cancer. *American Journal of Surgical Pathology*. 2020;44(9):1204-1212. doi:10.1097/PAS.0000000000001511
39. Mi EZ, Mi EZ, di Pietro M, et al. Comparative study of endoscopic surveillance in hereditary diffuse gastric cancer according to CDH1 mutation status. *Gastrointest Endosc*. 2018;87(2):408-418. doi:10.1016/j.gie.2017.06.028
40. van der Post RS, Gullo I, Oliveira C, et al. Histopathological, Molecular, and Genetic Profile of Hereditary Diffuse Gastric Cancer: Current Knowledge and Challenges for the Future. In: *Advances in Experimental Medicine and Biology*. Vol 908. ; 2016:371-391. doi:10.1007/978-3-319-41388-4_18
41. Riihimäki M, Hemminki A, Sundquist K, Sundquist J, Hemminki K. Metastatic spread in patients with gastric cancer. *Oncotarget*. 2016;7(32):52307-52316. doi:10.18632/oncotarget.10740
42. Rocha JP, Gullo I, Wen X, et al. Pathological features of total gastrectomy specimens from asymptomatic hereditary diffuse gastric cancer patients and implications for clinical management. *Histopathology*. 2018;73(6):878-886. doi:10.1111/his.13715
43. Choi IJ, Kim CG, Lee JY, et al. Family History of Gastric Cancer and *Helicobacter pylori* Treatment. *New England Journal of Medicine*. 2020;382(5):427-436. doi:10.1056/NEJMoa1909666

44. McCart Reed AE, Kalinowski L, Simpson PT, Lakhani SR. Invasive lobular carcinoma of the breast: the increasing importance of this special subtype. *Breast Cancer Research*. 2021;23(1):6. doi:10.1186/s13058-020-01384-6
45. Portschy PR, Marmor S, Nzara R, Virnig BA, Tuttle TM. Trends in Incidence and Management of Lobular Carcinoma In Situ: A Population-Based Analysis. *Ann Surg Oncol*. 2013;20(10):3240-3246. doi:10.1245/s10434-013-3121-4
46. Li CI, Anderson BO, Daling JR, Moe RE. Trends in Incidence Rates of Invasive Lobular and Ductal Breast Carcinoma. *JAMA*. 2003;289(11):1421. doi:10.1001/jama.289.11.1421
47. Thomas M, Kelly ED, Abraham J, Kruse M. Invasive lobular breast cancer: A review of pathogenesis, diagnosis, management, and future directions of early stage disease. *Semin Oncol*. 2019;46(2):121-132. doi:10.1053/j.seminoncol.2019.03.002
48. Van Baelen K, Geukens T, Maetens M, et al. Current and future diagnostic and treatment strategies for patients with invasive lobular breast cancer. *Annals of Oncology*. 2022;33(8):769-785. doi:10.1016/j.annonc.2022.05.006
49. Mejdahl MK, Wohlfahrt J, Holm M, et al. Synchronous bilateral breast cancer: a nationwide study on histopathology and etiology. *Breast Cancer Res Treat*. 2020;182(1):229-238. doi:10.1007/s10549-020-05689-0
50. Ansquer Y, Delaney S, Santulli P, Salomon L, Carbonne B, Salmon R. Risk of invasive breast cancer after lobular intra-epithelial neoplasia: Review of the literature. *European Journal of Surgical Oncology (EJSO)*. 2010;36(7):604-609. doi:10.1016/j.ejso.2010.05.019
51. Caldas C, Carneiro F, Lynch HT, et al. Familial gastric cancer: overview and guidelines for management. *J Med Genet*. 1999;36(12):873-880.
52. Garcia-Pelaez J, Barbosa-Matos R, Lobo S, et al. Genotype-first approach to identify associations between CDH1 germline variants and cancer phenotypes: a multicentre study by the European Reference Network on Genetic Tumour Risk Syndromes. *Lancet Oncol*. 2023;24(1):91-106. doi:10.1016/S1470-2045(22)00643-X
53. Gregory SN, Davis JL. CDH1 and hereditary diffuse gastric cancer: a narrative review. *Chin Clin Oncol*. 2023;12(3):25-25. doi:10.21037/cco-23-36
54. Kaurah P, MacMillan A, Boyd N, et al. Founder and Recurrent CDH1 Mutations in Families With Hereditary Diffuse Gastric Cancer. *JAMA*. 2007;297(21):2360. doi:10.1001/jama.297.21.2360
55. Kole C, Charalampakis N, Sakellariou S, et al. Hereditary Diffuse Gastric Cancer: A 2022 Update. *J Pers Med*. 2022;12(12):2032. doi:10.3390/jpm12122032
56. Caldas C, Carneiro F, Lynch HT, et al. Familial gastric cancer: overview and guidelines for management. *J Med Genet*. 1999;36(12):873-880.
57. Pharoah PDP, Guilford P, Caldas C. Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. *Gastroenterology*. 2001;121(6):1348-1353. doi:10.1053/gast.2001.29611
58. Hansford S, Kaurah P, Li-Chang H, et al. Hereditary Diffuse Gastric Cancer Syndrome: CDH1 Mutations and Beyond. *JAMA Oncol*. 2015;1(1):23. doi:10.1001/jamaoncol.2014.168

59. Roberts ME, Ranola JMO, Marshall ML, et al. Comparison of *CDH1* Penetrance Estimates in Clinically Ascertained Families vs Families Ascertained for Multiple Gastric Cancers. *JAMA Oncol*. 2019;5(9):1325. doi:10.1001/jamaoncol.2019.1208
60. Xicola RM, Li S, Rodriguez N, et al. Clinical features and cancer risk in families with pathogenic *CDH1* variants irrespective of clinical criteria. *J Med Genet*. 2019;56(12):838-843. doi:10.1136/jmedgenet-2019-105991
61. Vécsey-Semjén B, Becker KF, Sinski A, et al. Novel colon cancer cell lines leading to better understanding of the diversity of respective primary cancers. *Oncogene*. 2002;21(30):4646-4662. doi:10.1038/sj.onc.1205577
62. Oliveira C, Ferreira P, Nabais S, et al. E-Cadherin (CDH1) and p53 rather than SMAD4 and Caspase-10 germline mutations contribute to genetic predisposition in Portuguese gastric cancer patients. *Eur J Cancer*. 2004;40(12):1897-1903. doi:10.1016/j.ejca.2004.04.027
63. Barbosa-Matos R, Silva RL, Garrido L, et al. The cdh1 c.1901c>t variant: A founder variant in the portuguese population with severe impact in mrna splicing. *Cancers (Basel)*. 2021;13(17):1-18. doi:10.3390/cancers13174464
64. Gullo I, Devezas V, Baptista M, et al. Phenotypic heterogeneity of hereditary diffuse gastric cancer: report of a family with early-onset disease. *Gastrointest Endosc*. 2018;87(6):1566-1575. doi:10.1016/j.gie.2018.02.008
65. Evans JA. Old meets new: identifying founder mutations in genetic disease. *Can Med Assoc J*. 2015;187(2):93-94. doi:10.1503/cmaj.141509
66. Lee K, Krempely K, Roberts ME, et al. Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline CDH1 sequence variants. *Hum Mutat*. 2018;39(11):1553-1568. doi:10.1002/humu.23650
67. Flaum N, Crosbie EJ, Edmondson RJ, Smith MJ, Evans DG. Epithelial ovarian cancer risk: A review of the current genetic landscape. *Clin Genet*. 2020;97(1):54-63. doi:10.1111/cge.13566
68. Scalia-Wilbur J, Colins BL, Penson RT, Dizon DS. Breast Cancer Risk Assessment: Moving Beyond BRCA 1 and 2. *Semin Radiat Oncol*. 2016;26(1):3-8. doi:10.1016/j.semradonc.2015.09.004
69. Broca PP. *Traite Des Tumeurs*. Vol 1. P. Asselin; 1866.
70. Basset-Seguin N, Moles JP, Mils V, Dereure O, Guilhou JJ. TP53 tumor-suppressor gene and human carcinogenesis. *Exp Dermatol*. 1993;2(3):99-105. doi:10.1111/j.1600-0625.1993.tb00016.x
71. Rowell S, Newman B, Boyd J, King MC. Inherited predisposition to breast and ovarian cancer. *Am J Hum Genet*. 1994;55(5):861-865.
72. Wildrick DM. Molecular genetic studies of colon cancer. *Hematol Oncol Clin North Am*. 1989;3(1):1-18.
73. Lange K, Weeks D, Boehnke M, MacCluer JeanW, MacCluer JeanW. Programs for pedigree analysis: Mendel, Fisher, and dGene. *Genet Epidemiol*. 1988;5(6):471-472. doi:10.1002/gepi.1370050611
74. van der Post RS, Vogelaar IP, Carneiro F, et al. Hereditary diffuse gastric cancer: updated clinical guidelines with an emphasis on germline *CDH1* mutation carriers. *J Med Genet*. 2015;52(6):361-374. doi:10.1136/jmedgenet-2015-103094

75. Hankey BF, Ries LA, Edwards BK. The surveillance, epidemiology, and end results program: a national resource. *Cancer Epidemiol Biomarkers Prev*. 1999;8(12):1117-1121.
76. Coudert M, Drouet Y, Delhomelle H, et al. First estimates of diffuse gastric cancer risks for carriers of *CTNNA1* germline pathogenic variants. *J Med Genet*. 2022;59(12):1189-1195. doi:10.1136/jmg-2022-108740
77. Bonaïti B, Bonadona V, Perdry H, Andrieu N, Bonaïti-Pellié C. Estimating penetrance from multiple case families with predisposing mutations: extension of the 'genotype-restricted likelihood' (GRL) method. *European Journal of Human Genetics*. 2011;19(2):173-179. doi:10.1038/ejhg.2010.158
78. Oliveira C, Seruca R, Carneiro F. Hereditary gastric cancer. *Best Pract Res Clin Gastroenterol*. 2009;23(2):147-157. doi:10.1016/j.bpg.2009.02.003
79. de Klerk E, den Dunnen JT, 't Hoen PAC. RNA sequencing: from tag-based profiling to resolving complete transcript structure. *Cellular and Molecular Life Sciences*. 2014;71(18):3537-3551. doi:10.1007/s00018-014-1637-9
80. LeBlanc VG, Marra MA. Next-Generation Sequencing Approaches in Cancer: Where Have They Brought Us and Where Will They Take Us? *Cancers (Basel)*. 2015;7(3):1925-1958. doi:10.3390/cancers7030869
81. Umu SU, Langseth H, Keller A, et al. A 10-year prediagnostic follow-up study shows that serum RNA signals are highly dynamic in lung carcinogenesis. *Mol Oncol*. 2020;14(2):235-247. doi:10.1002/1878-0261.12620
82. Sorokin M, Ignatev K, Poddubskaya E, et al. RNA Sequencing in Comparison to Immunohistochemistry for Measuring Cancer Biomarkers in Breast Cancer and Lung Cancer Specimens. *Biomedicines*. 2020;8(5):114. doi:10.3390/biomedicines8050114
83. Horvath A, Pakala SB, Mudvari P, et al. Novel Insights into Breast Cancer Genetic Variance through RNA Sequencing. *Sci Rep*. 2013;3(1):2256. doi:10.1038/srep02256
84. Kothari V, Wei I, Shankar S, et al. Outlier Kinase Expression by RNA Sequencing as Targets for Precision Therapy. *Cancer Discov*. 2013;3(3):280-293. doi:10.1158/2159-8290.CD-12-0336
85. Hlevnjak M, Schulze M, Elgaafary S, et al. CATCH: A Prospective Precision Oncology Trial in Metastatic Breast Cancer. *JCO Precis Oncol*. 2021;(5):676-686. doi:10.1200/PO.20.00248
86. Tong H, Wang J, Chen H, Wang Z, Fan H, Ni Z. Transcriptomic analysis of gene expression profiles of stomach carcinoma reveal abnormal expression of mitotic components. *Life Sci*. 2017;170:41-49. doi:10.1016/j.lfs.2016.12.001
87. Carino A, Graziosi L, Marchianò S, et al. Analysis of Gastric Cancer Transcriptome Allows the Identification of Histotype Specific Molecular Signatures With Prognostic Potential. *Front Oncol*. 2021;11. doi:10.3389/fonc.2021.663771
88. Fitzgerald RC, Hardwick R, Huntsman D, et al. Hereditary diffuse gastric cancer: updated consensus guidelines for clinical management and directions for future research. *J Med Genet*. 2010;47(7):436-444. doi:10.1136/jmg.2009.074237

89.Corso G, Corso F, Bellerba F, et al. Geographical Distribution of E-cadherin Germline Mutations in the Context of Diffuse Gastric Cancer: A Systematic Review. *Cancers (Basel)*. 2021;13(6):1269. doi:10.3390/cancers13061269

90.National Center for Biotechnology Information. ClinVar; [VCV000012244.16], https://www.ncbi.nlm.nih.gov/clinvar/variation/VCV000012244.16 (accessed Sept. 14, 2023).

91.Oliveira C, Moreira H, Seruca R, Cardoso de Oliveira M, Carneiro F. Role of pathology in the identification of hereditary diffuse gastric cancer: report of a Portuguese family. *Virchows Archiv*. 2005;446(2):181-184. doi:10.1007/s00428-004-1156-4

92.Nam JH, Choi IJ, Cho SJ, et al. Helicobacter pylori infection and histological changes in siblings of young gastric cancer patients. *J Gastroenterol Hepatol*. 2011;26(7):1157-1163. doi:10.1111/j.1440-1746.2011.06717.x

93.El–Omar EM, Oien K, Murray LS, et al. Increased prevalence of precancerous changes in relatives of gastric cancer patients: Critical role of H. pylori. *Gastroenterology*. 2000;118(1):22-30. doi:10.1016/S0016-5085(00)70410-0

94.Shin CM, Kim N, Yang HJ, et al. Stomach Cancer Risk in Gastric Cancer Relatives. *J Clin Gastroenterol*. 2010;44(2):e34-e39. doi:10.1097/MCG.0b013e3181a159c4

95.Venneman K, Huybrechts I, Gunter MJ, Vandendaele L, Herrero R, Van Herck K. The epidemiology of *Helicobacter pylori* infection in Europe and the impact of lifestyle on its natural evolution toward stomach cancer after infection: A systematic review. *Helicobacter*. 2018;23(3):e12483. doi:10.1111/hel.12483

96.Bonadona V, Bonaïti B, Olschwang S, et al. Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in Lynch syndrome. *JAMA*. 2011;305(22):2304-2310. doi:10.1001/jama.2011.743

97.van der Post RS, Bisseling TM, van Dieren JM. Endoscopic surveillance: time for a paradigm shift in hereditary diffuse-type gastric cancer management? *Lancet Oncol*. 2023;24(4):311-312. doi:10.1016/S1470-2045(23)00094-3

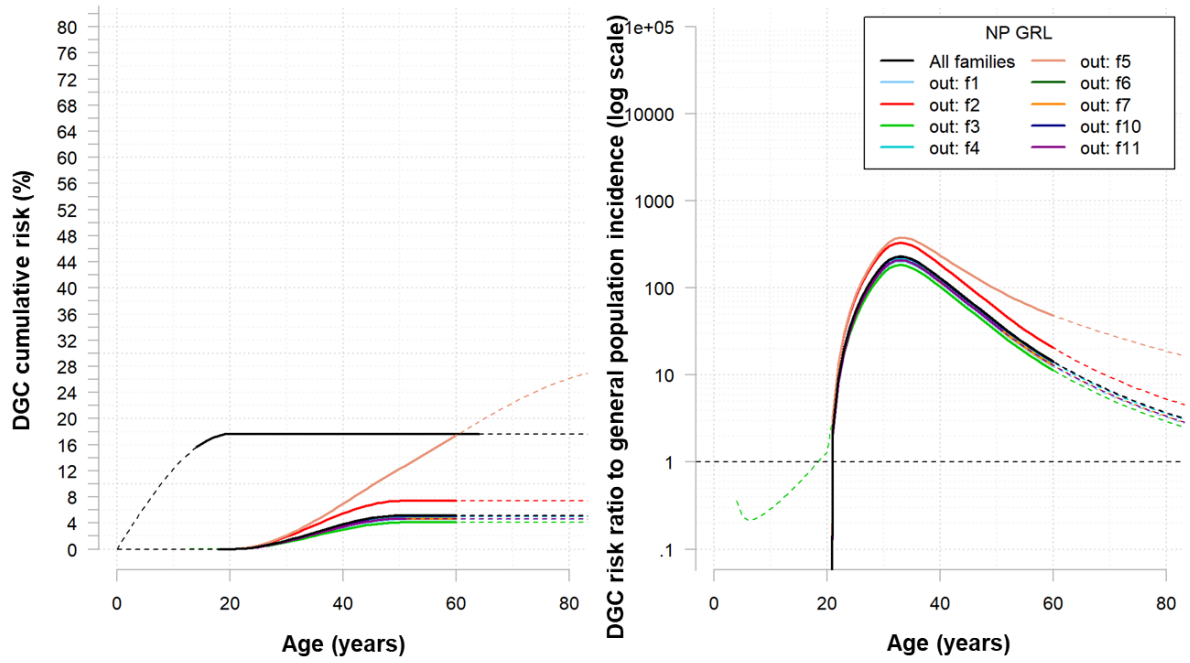
98.Gallanis AF, Davis JL. Unique challenges of risk-reducing surgery for hereditary diffuse gastric cancer syndrome: a narrative review. *European Journal of Cancer Prevention*. 2023;32(4):391-395. doi:10.1097/CEJ.0000000000000798

99.Jínek T, Adamčík L, Vrba R, Duda M, Škrovina M. Risk factors and post-operative complications after gastrectomy for cancer. *Rozhl Chir*. 2018;97(8):384-393.

9. Annexes

Supplementary Table 1. Standard Incidence Ratio for DGC (≥pT2 and pT1a)

DGC age of last news	c.1901C>T carriers (n = 95)	DGC cases* (n = 38)	SIR	SIR 95%CI inferior	SIR 95%CI superior
[0,35]	34	19	9234.12	5559.54	14420.23
(35,50]	31	11	1019.91	509.13	1824.90
(50,100]	30	11	75.83	32.74	149.42



Supplementary Figure 2. Leave-one-out sensitive analysis for risk estimations of DGC onset (with pT1a). **A.** Cumulative risk for DGC onset considering both pT1a and ≥pT2 tumours in black. A curve for every estimation leaving one of the families out of consideration. **B.** Risk ratio for DGC onset considering both pT1a and ≥pT2 tumours in HDGC against general population in black. A curve for every estimation leaving one of the families out of consideration.